AMENDMENTS TO THE DRAWINGS

New corrected drawings were required because Figs. 2, 3, and 6 were illegible. The drawings have been corrected to make Figs. 2, 3, and 6 legible. Corrected drawing sheets are being submitted with a separate letter to the Office Draftsman. The legend "REPLACEMENT SHEET" is included at the top of each sheet of corrected drawing sheets.

REMARKS/ARGUMENTS

The Office Action mailed January 4, 2007 has been carefully reviewed. Reconsideration of this application, as amended and in view of the following remarks, is respectfully requested. Claims 1-13 were originally in the application. Claim 8 stands withdrawn from consideration as being drawn to a non-elected species. The claims presented for examination are: claims 1-7 and 9-13.

Drawings

New corrected drawings were required because Figs. 2, 3, and 6 were illegible. The drawings have been corrected to make Figs. 2, 3, and 6 legible. Corrected drawing sheets are being submitted with a separate letter to the Office Draftsman. The legend "REPLACEMENT SHEET" is included at the top of each sheet of corrected drawing sheets.

The corrected Figs. 2, 3, and 6 are the best black and white versions of the figures that can be made. The original color versions of Figs. 2, 3, and 6 appear in the article LGA: a method for finding 3D similarities in protein structures by Adam Zemla, in the journal, Nucleic Acids Research, 2003, Vol. 31, No. 13, pp. 3370-3374. The specification has been amended to include citation to the article.

Fig. 2 which shows a strip chart used to plot output from the standard structure comparison analysis of protein structures is Figure 1A in the article. Fig. 3 which shows the strip chart representing the results from the LGA is Figure 1B in the article. Fig. 6 is a bar representation of the results from sequence independent LGA superpositions is Figure 2A in the article. A copy of the article, "LGA: a method for finding 3D similarities in protein structures" by Adam Zemla, in the journal, Nucleic Acids Research, 2003, Vol. 31, No. 13, pp. 3370-

3374 is enclosed. Applicants believe that the amendment overcomes the objection the drawings and that a complete response to the rejection has been provided.

Claim Objections

Claims 1 and 6 were objected to because in claim 1 the phrase in the preamble "finding 3D similarities in protein the structure" is not clear and claim 6 consists of two sentences. With regard to claim 1 the Office Action stated, "it seems that 'finding 3D similarities in protein structures of' is rather meant."

Applicants have amended claim 1 and 6 to overcome the objections. Claim 1 has been amended to delete the word "the" so that the phrase reads "finding 3D similarities in protein structures of." Claim 6 has been amended to be a single sentence. Applicants believe that the amendment overcomes the objection to claim 1 and 6 and that a complete response to the rejection has been provided.

Claim Rejections - 35 USC § 112, Second Paragraph

Claims 1-7 and 9-13 were rejected under 35 U.S.C. § 112, second paragraph, as being indefinite for failing to particularly point out and distinctly claim the subject matter which applicant regards as the invention. Applicants will respond to each portion of this rejection using the nomenclature (#A, #B, #C, etc.) used in the Office Action.

#A. Claim 1

The Office Action stated, "It is not clear whether the three steps of 'comparing' recited in the claim are conducted independently or sequentially."

Applicants have amended claim to clarify that the three steps of "comparing" are conducted independently.

#B. Claim 1

The Office Action stated, "The step of using LGA_S analysis is vague and indefinite: specification, in the section defining the scoring function LGA_S (p. 17), fails to identify variables k, X, S(F), S(LCS),w* used in calculation of the scoring function; thus, the function is undefined."

Applicants have amended paragraph [0036] of the specification to indicate that, "The Local Global Alignment Scoring function (LGA_S) is described in greater detail in the following portions of the DETAILED DESCRIPTION OF THE INVENTION, particularly in paragraph [0036]. The journal article LGA: a method for finding 3D similarities in protein structures by Adam Zemla, in the journal, Nucleic Acids Research, 2003, Vol. 31, No. 13, pp. 3370-3374 includes a section under the heading 'Description of the LGA scoring function' which provides additional information about using Local Global Alignment Scoring function (LGA_S). The journal article LGA: a method for finding 3D similarities in protein structures by Adam Zemla, in the journal, Nucleic Acids Research, 2003, Vol. 31, No. 13, pp. 3370-3374 is incorporated herein in its entirety by this reference." Applicants believe the specification and the journal article LGA: a method for finding 3D similarities in protein structures by Adam Zemla describe the Local Global Alignment Scoring function (LGA_S) sufficiently for an individual skilled in the art to practice the claimed invention and that the claim step of using LGA_S analysis complies with 35 U.S.C. § 112, second paragraph.

#C. Claims 3 and 9

The Office Action stated, "Similarly, the term 'apply the transform' in claims 3, 9 is vague and indefinite. The term 'transform' is not defined by the claim. The specification, also, does not provide a standard for ascertaining the requisite composition, and one of ordinary skills in the art would not be reasonably appraised of the scope of the invention."

Applicants believe the term "apply the transform" in claims 3 and 9 sufficiently for an individual skilled in the art to practice the claimed invention and that claims 3 and 9 comply with 35 U.S.C. § 112, second paragraph.

#D. Claim 1

The Office Action stated, "claim 1, last step: it is not clear what constitutes 'the calculated alignment.' The preceding steps of the claim are based on preexisting 'structure information of alignment' (first step of the claim) and no method step is directed to calculating alignment. Further, it is not clear what is meant by 'quality' of the calculated alignment, and what constitutes 'complete information' about the quality of the alignment."

Applicants have amended claim 1 to clarify that the step of, "independently comparing the first molecule and the second molecule using said preselected information and using Local Global Alignment Scoring function (LGA_S) analysis to provide a calculated alignment between said first molecule and said second molecule" and to delete the qualitative words "complete" and "quality."

#E. Claim 3, step b)

The Office Action stated, "#E. Claim 3, step b): the claim step contains two alternative steps, 'verify' and 'modify' an alignment. A broad range or limitation together with a narrow range or limitation that falls within the broad range or limitation (in the same claim) is considered indefinite, since the resulting claim does not clearly set forth the metes and bounds of the patent protection desired. See MPEP § 2173.05(c). The Board stated that this can render a claim indefinite by raising a question or doubt as to whether the feature introduced by such language is (a) merely exemplary of the remainder of the claim, and therefore not required, or (b) a required feature of the claims. Note also, for example, the decisions of Ex We Steigewald, 131 USPQ 74 (Bd. App. 1961); Ex parte Hail, 83

USPQ 38 (Bd. App. 1948); and Ex pane Hasche, 86 USPQ 481 (Bd. App. 1949). In the present instance, claim 3 recites the broad recitation 'modify,' and the claim also recites 'verify' which is the narrower statement of the limitation."

Applicants have amended claim 3 to remove one of the two alternative steps. The alternative step "verify" has been deleted.

#F. Claim 7

The Office Action stated, "The term 'preselected information' is not clear. There is no antecedent basis for the term - there is no 'preselected information addressed in the preceding part of the claim.'"

Applicants have amended claim 7 to include the step, "preselecting structure information of alignment of residue-residue correspondence for said first molecule and said second molecule to produce preselected information." The addition of this step provides and antecedent basis for the subsequent term "preselected information."

Rejections Overcome

Applicants believe that the amendments and explanations overcome the rejection of claims 1-7 and 9-13 under 35 U.S.C. § 112, second paragraph, and that a complete response to the rejection has been provided.

Claim Rejections - 35 U.S.C. § 101 (utility)

Claims 1-7 and 9-13 were rejected under 35 U.S.C. § 101. The Office Action mailed January 4, 2007 alleged that the claimed invention lacks utility.

Applicants respectfully traverse this rejection. Applicants' claimed invention has many worthwhile uses and has utility. For example, Applicants' claimed invention provides information about protein folding that furthers research into finding a cure for diseases such as Alzheimer's disease, Mad Cow disease, Amyotrophic Lateral Sclerosis (ALS), Huntington's disease, Parkinson's

disease, and many Cancers. The 2000-2006 article by Vijay Pande and Stanford University, http://folding.stanford.edu/, states, "Proteins are biology's workhorses -- its 'nanomachines.' Before proteins can carry out these important functions, they assemble themselves, or 'fold.' The process of protein folding, while critical and fundamental to virtually all of biology, in many ways remains a mystery. Moreover, when proteins do not fold correctly (i.e., 'misfold'), there can be serious consequences, including many well known diseases, such as Alzheimer's, Mad Cow (BSE), CJD, ALS, Huntington's, Parkinson's disease, and many Cancers and cancer-related syndromes." A copy of the 2000-2006 article by Vijay Pande and Stanford University is enclosed.

Applicants' claimed invention allows the comparison of 3D similarities in protein structure of a first molecule and of a second molecule. This has utility, for example if there is a substantial body of knowledge about the first molecule then the information obtained about the comparison provides information about the second molecule.

In addition, evidence that others have used the invention and have commented favorably on the invention shows that the invention has utility. There are an extensive number of examples of others having used the invention in studying diseases and in developing new pharmaceuticals. Applicants will discuss some examples of these uses; however, the fact that there are extensive uses of the invention demonstrates that that the invention has utility.

The journal article "Consensus sequences improve PSI-BLAST through mimicking profile–profile alignments" by Dariusz Przybylski and Burkhard Rost; in Nucleic Acids Research, 2007, Vol. 35, No. 7, pp. 2238–2246, published online March 16, 2007, under the heading "Evaluation of Alignment Quality" contains the following paragraph:

"We superposed all models (represented by Ca atom coordinates) with experimentally determined 3D structures using one particular automatic method for structural superposition, namely LGA (40); this method has become one of the standards in the experiments for the Critical Assessment of Structure Prediction [CASP (41)]. First, we computed a Global Distance Test (GDT) (40) that corresponds to the largest, not necessarily continuous subset of residues superimposable within a specified distance threshold. Second, we also computed Longest Continuous Segments (LCS) (40) of residues (consecutively modeled residues) that can fit under a specified RMSD cutoff. The second measure provided us with a local alignment quality test. Note that we chose a subset of pairs (Q,T) such that for all pairs experimental structures were available; we built the model for Q using the known structure of T and assuming that Q had no known structure, but we evaluated the accuracy of the model using the experimentally known structure for Q. We reported results for two different thresholds. The first was rather stringent (2Å); it focused on the essential core similarities between model and experiment. The second was rather relaxed (5Å) thereby capturing more generic, coarse-grained similarities. Note that GDT computation uses the actual distance threshold while LCS uses average distance (RMSD)."

The fact that the authors Dariusz Przybylski and Burkhard Rost state in the article, ".... one particular automatic method for structural superposition, namely LGA; this method has become one of the standards in the experiments for the Critical Assessment of Structure Prediction" and that the authors have used the invention, is evidence that Applicants' specification which includes the article shows that the invention has utility. A copy of the article, "Consensus sequences improve PSI-BLAST through mimicking profile–profile alignments" by Dariusz Przybylski and Burkhard Rost; in Nucleic Acids Research, 2007, Vol. 35, No. 7, pp. 2238–2246, published online March 16, 2007, is enclosed.

The journal article "Associative memory Hamiltonians for structure prediction without homology: α/β proteins" by Corey Hardin, Michael P. Eastwood, Michael C. Prentiss, Zadia Luthey-Schulten, and Peter G. Wolynes; in Proceedings of the National Academy of Sciences of the United States of

America, published online before print January 28, 2003, 10.1073/pnas.252753899, contains the following paragraph:

> "The global superposition of two structures can often fail to highlight significant segments of correct native structure. We have submitted the best Q structures to the LGA (Local-Global Alignment) server (http://PredictionCenter.llnl.gov/local/lga), and the results are given in Table 6. The predictions are evaluated according to two measures, LCS and GDT. LCS is the longest continuous (along the sequence) segment that can be superimposed on the native structure without exceeding a rmsd cutoff. The global distance test (GDT) represents the largest number of residues that lie within a distance cutoff of their correct positions. The set of residues need not be contiguous. In all three cases large portions of the prediction are correct to within the cutoff. We have also used CE to align the predicted and native structure. Note that in all three cases the predicted structure is more similar to the native than any of the database structures, thus demonstrating the ability of the potential to generalize from incorrect scaffolds. The scores of local similarity will, of course, depend on the chosen cutoff. Fig. 2 is a Hubbard plot of the percent of residues below the cutoff, as a function of the cutoff distance."

The fact that the authors Corey Hardin, Michael P. Eastwood, Michael C. Prentiss, Zadia Luthey-Schulten, and Peter G. Wolynes have used the invention, stated in the article as follows: "We have submitted the best Q structures to the LGA (Local-Global Alignment) server (http://PredictionCenter.llnl.gov/local/lga), and the results are given in Table 6," is evidence that the specification which includes the article shows that the invention has utility. A copy of the article, "Associative memory Hamiltonians for structure prediction without homology: α/β proteins" by Corey Hardin, Michael P. Eastwood, Michael C. Prentiss, Zadia Luthey-Schulten, and Peter G. Wolynes; in *Proceedings of the National Academy of Sciences of the United States of America*, published online before print January 28, 2003, 10.1073/pnas.252753899, is enclosed.

The journal article "Crystal Structure of *Mycobacterium tuberculosis* Diaminopimelate Decarboxylase, an Essential Enzyme in Bacterial Lysine

Biosynthesis" by Kuppan Gokulan, Bernhard Rupp, Martin S. Pavelka Jr., William R. Jacobs Jr., and James C. Sacchettini; in J. Biol. Chem., Vol. 278, Issue 20, 18588-18596, May 16, 2003, contains the following paragraph:

> "Fig. 4. Backbone superposition of known fold type III PLP-dependent enzyme structures. Panel A, color key: cyan, M. tuberculosis DAPDC; magenta, human ODC; green, mouse ODC; and yellow, T. brucei. Panel B, superposition of M. tuberculosis DAPDC (cyan) with B. stearothermopilus AR (red). The rotation of the AR β-domain relative to the other structures is clearly visible. The superpositions were carried by the Local-Global-Alignment server (Adam Zemla, predictioncenter.llnl.gov/local/lga/lga.html); corresponding r.m.s.d. values are listed in Table III."

The fact that the authors Kuppan Gokulan, Bernhard Rupp, Martin S. Pavelka Jr., William R. Jacobs Jr., and James C. Sacchettini have used the invention, stated in the article as follows: "The superpositions were carried by the Local-Global-Alignment server (Adam Zemla, predictioncenter.llnl.gov/local/lga/lga.html); corresponding r.m.s.d. values are listed in Table III," is evidence that the specification which includes the article shows that the invention has utility. A copy of the article, "Crystal Structure of *Mycobacterium tuberculosis* Diaminopimelate Decarboxylase, an Essential Enzyme in Bacterial Lysine Biosynthesis" by Kuppan Gokulan, Bernhard Rupp, Martin S. Pavelka Jr., William R. Jacobs Jr., and James C. Sacchettini; in J. Biol. Chem., Vol. 278, Issue 20, 18588-18596, May 16, 2003, is enclosed.

The journal article "Structural and Functional Basis of CXCL12 (Stromal Cell-derived Factor-1α) Binding to Heparin" by James W. Murphy, Yoonsang Cho, Aristidis Sachpatzidis, Chengpeng Fan, Michael E. Hodsdon, and Elias Lolis; in THE JOURNAL OF BIOLOGICAL CHEMISTRY VOL. 282, NO. 13, pp. 10018–10027, March 30, 2007, contains the following paragraph:

> "The crystal structure of the complex is shown in Fig. 5A. As in the native structure three ß-strands from each subunit form an extended six-stranded ß-sheet that in turn forms a pocket where a heparin disaccharide molecule is

located. Root mean square deviation values were determined for each α carbon between the apo and bound structures using a local global alignment similarity calculation (44). 44. Zemla, A. (2003) Nucleic Acids Res. 31, 3370–3374."

The fact that the authors James W. Murphy, Yoonsang Cho, Aristidis Sachpatzidis, Chengpeng Fan, Michael E. Hodsdon, and Elias Lolis have used the invention, stated in the article as follows: in the article, "....using a local global alignment similarity calculation (44). 44. Zemla, A. (2003) Nucleic Acids Res. 31, 3370–3374," is evidence that Applicants' specification which includes the article shows that the invention has utility. A copy of the article, "Structural and Functional Basis of CXCL12 (Stromal Cell-derived Factor-1α) Binding to Heparin" by James W. Murphy, Yoonsang Cho, Aristidis Sachpatzidis, Chengpeng Fan, Michael E. Hodsdon, and Elias Lolis; in THE JOURNAL OF BIOLOGICAL CHEMISTRY VOL. 282, NO. 13, pp. 10018–10027, March 30, 2007," is enclosed.

The journal article "On the Accuracy of Homology Modeling and Sequence Alignment Methods Applied to Membrane Proteins" by Lucy R. Forrest, Christopher L. Tang, and Barry Honig; in Biophysical Journal, Volume 91, July 2006, pp. 508–517, contains the following paragraph:

"Gap penalties were also assigned according to the location of core regions or secondary-structure elements. We used the local-global alignment method where unaligned terminal residues are only penalized in the query. 39. Zemla, A. 2003. LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Res. 31:3370–3374."

The fact that the authors Lucy R. Forrest, Christopher L. Tang, and Barry Honi have used the invention, stated in the article as follows: "We used the local-global alignment method ....," is evidence that Applicants' specification which includes the article shows that the invention has utility. A copy of the article, "On the Accuracy of Homology Modeling and Sequence Alignment Methods Applied

to Membrane Proteins" by Lucy R. Forrest, Christopher L. Tang, and Barry Honig; in Biophysical Journal, Volume 91, July 2006, pp. 508–517," is enclosed.

The journal article "Antibody Elbow Angles are Influenced by their Light Chain Class" by Robyn L. Stanfield, Adam Zemla, Ian A. Wilson, and Bernhard Rupp; in J. Mol. Biol. (2006) 357, 1566–1574, contains the following paragraph:

> "The program RBOW is implemented on a Linux platform (Apache web server) using a combination of Perl scripts and standard Fortran90 code. The alignment program used is the local-global alignment program LGA, which assures minimal dependence of the results on the Fab numbering convention used. The program allows upload of PDB format coordinate files, selection of heavy and light chain identifiers, and input of domain boundaries, for which reasonable default values are provided."

The fact that the authors Robyn L. Stanfield, Adam Zemla, Ian A. Wilson, and Bernhard Rupp have used the invention, stated in the article as follows: "The alignment program used is the local-global alignment program LGA," is evidence that Applicants' specification which includes the article shows that the invention has utility. A copy of the article, "Antibody Elbow Angles are Influenced by their Light Chain Class" by Robyn L. Stanfield, Adam Zemla, Ian A. Wilson, and Bernhard Rupp; in J. Mol. Biol. (2006) 357, 1566–1574," is enclosed.

The journal article "Mycobacterium tuberculosis RmlC epimerase (Rv3465): a promising drug-target structure in the rhamnose pathway " by Katherine A. Kantardjieff, Chang-Yub Kim, Cleo Naranjo, Geoffry S. Waldo, Timothy Lekin, Brent W. Segelke, Adam Zemla, Min S. Park, Thomas C. Terwilliger, and Bernhard Rupp; in Acta Cryst. (2004). D60, 895–902, contains the following paragraph:

"Figure 1 - Pairwise structural alignment of homologous protein chains with RmlC from MTB using the local global alignment program LGA (Zemla, 2003)."

The fact that the authors Katherine A. Kantardjieff, Chang-Yub Kim, Cleo Naranjo, Geoffry S.Waldo, Timothy Lekin, Brent W. Segelke, Adam Zemla, Min S. Park, Thomas C. Terwilliger, and Bernhard Rupp have used the invention, stated in the article as follows: ".... using the local global alignment program LGA (Zemla, 2003)," is evidence that Applicants' specification which includes the article shows that the invention has utility. A copy of the article, "Mycobacterium tuberculosis RmlC epimerase (Rv3465): a promising drug-target structure in the rhamnose pathway" by Katherine A. Kantardjieff, Chang-Yub Kim, Cleo Naranjo, Geoffry S.Waldo, Timothy Lekin, Brent W. Segelke, Adam Zemla, Min S. Park, Thomas C. Terwilliger, and Bernhard Rupp; in Acta Cryst. (2004). D60, 895–902 is enclosed.

Claim Rejections - 35 USC § 112, First Paragraph.

Claims 1-7 and 9-13 were rejected under 35 U.S.C. § 112, first paragraph. The Office Action mailed January 4, 2007 stated, "Specifically, since the claimed invention is not supported by either a credible asserted utility or a well established utility, one skilled in the art would not know how to use the claimed invention. In addition, due to ambiguity of the term 'to transform,' and variables used in determining the LGA_S score (see rejections under 35 U.S.C. § 112, second paragraph, sections B)-C)), the specification is not enabling as one skilled in the art would not know how to make, and thus how to use, the invention as claimed."

Applicants submit that the specification contains details sufficient for one skilled the art to make and use the invention defined by claims 1-7 and 9-13. Not

everything necessary to practice the invention need be disclosed. In re Buchner, 929 F.2d 660, 661 (Fed. Cir. 1991). Applicants submit that the specification as originally filed enables the practice of the invention defined by claims 1-7 and 9-13. Applicants submit (1) there was considerable direction and guidance in the specification, (2) there was a high level of skill in the art at the time the application was filed, and (3) all of the background needed to practice the invention was known.

The journal article "LGA - a method for finding 3D similarities in protein structures," by Adam Zemla, Nucleic Acids Research, Vol. 31, No. 13, 2003, pp. 3370-3374, is incorporated into the subject patent application by reference. The fact that the article by the inventor Adam Zemla was reviewed and accepted by a major journal is evidence that the specification which includes the article contains details sufficient for one skilled the art to make and use the invention defined by claims 1-7 and 9-13.

In addition, evidence that others have used the invention and have commented favorably on the invention establishes that Applicants' specification (including the incorporated article) contains details sufficient for one skilled the art to make and use the invention defined by claims 1-7 and 9-13. There are an extensive number of examples of others having used the invention in studying diseases and in developing new pharmaceuticals. Applicants will discuss some examples of these uses; however, the fact that there are extensive uses of the invention demonstrates that Applicants specification (including the incorporated article) contains details sufficient for one skilled the art to make and use the invention defined by claims 1-7 and 9-13.

The journal article "Consensus sequences improve PSI-BLAST through mimicking profile–profile alignments" by Dariusz Przybylski and Burkhard Rost; in Nucleic Acids Research, 2007, Vol. 35, No. 7, pp. 2238–2246, published

online March 16, 2007, under the heading "Evaluation of Alignment Quality" contains the following paragraph:

> "We superposed all models (represented by Ca atom coordinates) with experimentally determined 3D structures using one particular automatic method for structural superposition, namely LGA (40); this method has become one of the standards in the experiments for the Critical Assessment of Structure Prediction [CASP (41)]. First, we computed a Global Distance Test (GDT) (40) that corresponds to the largest, not necessarily continuous subset of residues superimposable within a specified distance threshold. Second, we also computed Longest Continuous Segments (LCS) (40) of residues (consecutively modeled residues) that can fit under a specified RMSD cutoff. The second measure provided us with a local alignment quality test. Note that we chose a subset of pairs (Q,T) such that for all pairs experimental structures were available; we built the model for Q using the known structure of T and assuming that Q had no known structure, but we evaluated the accuracy of the model using the experimentally known structure for Q. We reported results for two different thresholds. The first was rather stringent (2Å); it focused on the essential core similarities between model and experiment. The second was rather relaxed (5Å) thereby capturing more generic, coarse-grained similarities. Note that GDT computation uses the actual distance threshold while LCS uses average distance (RMSD)."

The fact that the authors Dariusz Przybylski and Burkhard Rost state in the article, ".... one particular automatic method for structural superposition, namely LGA; this method has become one of the standards in the experiments for the Critical Assessment of Structure Prediction" and that the authors have used the invention, is evidence that Applicants' specification which includes the article contains details sufficient for one skilled the art to make and use the invention defined by claims 1-7 and 9-13. A copy of the article, "Consensus sequences improve PSI-BLAST through mimicking profile–profile alignments" by Dariusz Przybylski and Burkhard Rost; in Nucleic Acids Research, 2007, Vol. 35, No. 7, pp. 2238–2246, published online March 16, 2007, is enclosed.

The journal article "Associative memory Hamiltonians for structure prediction without homology: α/β proteins" by Corey Hardin, Michael P.

Eastwood, Michael C. Prentiss, Zadia Luthey-Schulten, and Peter G. Wolynes; in Proceedings of the National Academy of Sciences of the United States of America, published online before print January 28, 2003, 10.1073/pnas.252753899, contains the following paragraph:

> "The global superposition of two structures can often fail to highlight significant segments of correct native structure. We have submitted the best Q structures to the LGA (Local-Global Alignment) server (http://PredictionCenter.llnl.gov/local/lga), and the results are given in Table 6. The predictions are evaluated according to two measures, LCS and GDT. LCS is the longest continuous (along the sequence) segment that can be superimposed on the native structure without exceeding a rmsd cutoff. The global distance test (GDT) represents the largest number of residues that lie within a distance cutoff of their correct positions. The set of residues need not be contiguous. In all three cases large portions of the prediction are correct to within the cutoff. We have also used CE to align the predicted and native structure. Note that in all three cases the predicted structure is more similar to the native than any of the database structures, thus demonstrating the ability of the potential to generalize from incorrect scaffolds. The scores of local similarity will, of course, depend on the chosen cutoff. Fig. 2 is a Hubbard plot of the percent of residues below the cutoff, as a function of the cutoff distance."

The fact that the authors Corey Hardin, Michael P. Eastwood, Michael C. Prentiss, Zadia Luthey-Schulten, and Peter G. Wolynes have used the invention, stated in the article as follows: "We have submitted the best Q structures to the LGA (Local-Global Alignment) server (http://PredictionCenter.llnl.gov/local/lga), and the results are given in Table 6," is evidence that the specification which includes the article contains details sufficient for one skilled the art to make and use the invention defined by claims 1-7 and 9-13. A copy of the article, "Associative memory Hamiltonians for structure prediction without homology: α/β proteins" by Corey Hardin, Michael P. Eastwood, Michael C. Prentiss, Zadia Luthey-Schulten, and Peter G. Wolynes; in *Proceedings of the National Academy of Sciences*

of the United States of America, published online before print January 28, 2003, 10.1073/pnas.252753899, is enclosed.

The journal article "Crystal Structure of *Mycobacterium tuberculosis* Diaminopimelate Decarboxylase, an Essential Enzyme in Bacterial Lysine Biosynthesis" by Kuppan Gokulan, Bernhard Rupp, Martin S. Pavelka Jr., William R. Jacobs Jr., and James C. Sacchettini; in J. Biol. Chem., Vol. 278, Issue 20, 18588-18596, May 16, 2003, contains the following paragraph:

> "Fig. 4. Backbone superposition of known fold type III PLP-dependent enzyme structures. Panel A, color key: cyan, M. tuberculosis DAPDC; magenta, human ODC; green, mouse ODC; and yellow, T. brucei. Panel B, superposition of M. tuberculosis DAPDC (cyan) with B. stearothermopilus AR (red). The rotation of the AR β-domain relative to the other structures is clearly visible. The superpositions were carried by the Local-Global-Alignment server (Adam Zemla, predictioncenter.llnl.gov/local/lga/lga.html); corresponding r.m.s.d. values are listed in Table III."

The fact that the authors Kuppan Gokulan, Bernhard Rupp, Martin S. Pavelka Jr., William R. Jacobs Jr., and James C. Sacchettini have used the invention, stated in the article as follows: The superpositions were carried by the Local-Global-Alignment server (Adam Zemla, predictioncenter.llnl.gov/local/lga/lga.html); corresponding r.m.s.d. values are listed in Table III, " is evidence that the specification which includes the article contains details sufficient for one skilled the art to make and use the invention defined by claims 1-7 and 9-13. A copy of the article, "Crystal Structure of *Mycobacterium tuberculosis* Diaminopimelate Decarboxylase, an Essential Enzyme in Bacterial Lysine Biosynthesis" by Kuppan Gokulan, Bernhard Rupp, Martin S. Pavelka Jr., William R. Jacobs Jr., and James C. Sacchettini; in J. Biol. Chem., Vol. 278, Issue 20, 18588-18596, May 16, 2003, is enclosed.

The journal article "Structural and Functional Basis of CXCL12 (Stromal Cell-derived Factor-1α) Binding to Heparin" by James W. Murphy, Yoonsang

Cho, Aristidis Sachpatzidis, Chengpeng Fan, Michael E. Hodsdon, and Elias Lolis; in THE JOURNAL OF BIOLOGICAL CHEMISTRY VOL. 282, NO. 13, pp. 10018–10027, March 30, 2007, contains the following paragraph:

> "The crystal structure of the complex is shown in Fig. 5A. As in the native structure three ß-strands from each subunit form an extended six-stranded ß-sheet that in turn forms a pocket where a heparin disaccharide molecule is located. Root mean square deviation values were determined for each α carbon between the apo and bound structures using a local global alignment similarity calculation (44). 44. Zemla, A. (2003) Nucleic Acids Res. 31, 3370–3374."

The fact that the authors James W. Murphy, Yoonsang Cho, Aristidis Sachpatzidis, Chengpeng Fan, Michael E. Hodsdon, and Elias Lolis have used the invention, stated in the article as follows: in the article, "....using a local global alignment similarity calculation (44). 44. Zemla, A. (2003) Nucleic Acids Res. 31, 3370–3374," is evidence that Applicants' specification which includes the article contains details sufficient for one skilled the art to make and use the invention defined by claims 1-7 and 9-13. A copy of the article, "Structural and Functional Basis of CXCL12 (Stromal Cell-derived Factor-1α) Binding to Heparin" by James W. Murphy, Yoonsang Cho, Aristidis Sachpatzidis, Chengpeng Fan, Michael E. Hodsdon, and Elias Lolis; in THE JOURNAL OF BIOLOGICAL CHEMISTRY VOL. 282, NO. 13, pp. 10018–10027, March 30, 2007," is enclosed.

The journal article "On the Accuracy of Homology Modeling and Sequence Alignment Methods Applied to Membrane Proteins" by Lucy R. Forrest, Christopher L. Tang, and Barry Honig; in Biophysical Journal, Volume 91, July 2006, pp. 508–517, contains the following paragraph:

> "Gap penalties were also assigned according to the location of core regions or secondary-structure elements. We used the local-global alignment method where unaligned terminal residues are only penalized in the query. 39. Zemla, A. 2003. LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Res. 31:3370–3374."

The fact that the authors Lucy R. Forrest, Christopher L. Tang, and Barry Honi have used the invention, stated in the article as follows: "We used the local-global alignment method ...," is evidence that Applicants' specification which includes the article contains details sufficient for one skilled the art to make and use the invention defined by claims 1-7 and 9-13. A copy of the article, "On the Accuracy of Homology Modeling and Sequence Alignment Methods Applied to Membrane Proteins" by Lucy R. Forrest, Christopher L. Tang, and Barry Honig; in Biophysical Journal, Volume 91, July 2006, pp. 508–517, is enclosed.

The journal article "Antibody Elbow Angles are Influenced by their Light Chain Class" by Robyn L. Stanfield, Adam Zemla, Ian A. Wilson, and Bernhard Rupp; in J. Mol. Biol. (2006) 357, 1566–1574, contains the following paragraph:

> "The program RBOW is implemented on a Linux platform (Apache web server) using a combination of Perl scripts and standard Fortran90 code. The alignment program used is the local-global alignment program LGA, which assures minimal dependence of the results on the Fab numbering convention used. The program allows upload of PDB format coordinate files, selection of heavy and light chain identifiers, and input of domain boundaries, for which reasonable default values are provided."

The fact that the authors Robyn L. Stanfield, Adam Zemla, Ian A. Wilson, and Bernhard Rupp have used the invention, stated in the article as follows: "The alignment program used is the local-global alignment program LGA," is evidence that Applicants' specification which includes the article contains details sufficient for one skilled the art to make and use the invention defined by claims 1-7 and 9-13. A copy of the article, "Antibody Elbow Angles are Influenced by their Light Chain Class" by Robyn L. Stanfield, Adam Zemla, Ian A. Wilson, and Bernhard Rupp; in J. Mol. Biol. (2006) 357, 1566–1574, is enclosed.

The journal article "Mycobacterium tuberculosis RmlC epimerase (Rv3465): a promising drug-target structure in the rhamnose pathway" by Katherine A. Kantardjieff, Chang-Yub Kim, Cleo Naranjo, Geoffry S.Waldo, Timothy Lekin, Brent W. Segelke, Adam Zemla, Min S. Park, Thomas C. Terwilliger, and Bernhard Rupp; in Acta Cryst. (2004). D60, 895–902, contains the following paragraph:

> "Figure 1 - Pairwise structural alignment of homologous protein chains with RmlC from MTB using the local global alignment program LGA (Zemla, 2003)."

The fact that the authors Katherine A. Kantardjieff, Chang-Yub Kim, Cleo Naranjo, Geoffry S.Waldo, Timothy Lekin, Brent W. Segelke, Adam Zemla, Min S. Park, Thomas C. Terwilliger, and Bernhard Rupp have used the invention, stated in the article as follows: ".... using the local global alignment program LGA (Zemla, 2003)," is evidence that Applicants' specification which includes the article contains details sufficient for one skilled the art to make and use the invention defined by claims 1-7 and 9-13. A copy of the article, "Mycobacterium tuberculosis RmlC epimerase (Rv3465): a promising drug-target structure in the rhamnose pathway" by Katherine A. Kantardjieff, Chang-Yub Kim, Cleo Naranjo, Geoffry S.Waldo, Timothy Lekin, Brent W. Segelke, Adam Zemla, Min S. Park, Thomas C. Terwilliger, and Bernhard Rupp; in Acta Cryst. (2004). D60, 895–902 is enclosed.

Applicants submit that the term "apply the transform" used in claims 3 and 9 is described in Applicants original specification sufficiently for an individual skilled in the art to practice the claimed invention and that claims 3 and 9 comply with 35 U.S.C. § 112, first paragraph. Applicants also submit that the variables used in determining the LGA_S score described in Applicants original specification and in the journal article LGA: a method for finding 3D

similarities in protein structures by Adam Zemla, in the journal, Nucleic Acids Research, 2003, Vol. 31, No. 13, pp. 3370-3374 incorporated into the specification by reference sufficiently for an individual skilled in the art to practice the claimed invention and that claims 3 and 9 comply with 35 U.S.C. § 112, first paragraph.

Claim Rejections - 35 U.S.C. § 101 (non-statutory invention)

Claims 1-7 and 9-13 were rejected under 35 U.S.C. § 101. The Office Action mailed January 4, 2007 alleged that the claimed invention is directed to non-statutory subject matter. The Office Action mailed January 4, 2007 stated that, "the invention does not satisfy the criteria of utility requirements as not being substantial."

Applicants respectfully traverse this rejection. Applicants' claimed invention is "substantial." Applicants' claimed invention has many worthwhile uses. For example, Applicants' claimed invention provides information about protein folding that furthers research into finding a cure for diseases such as Alzheimer's disease, Mad Cow disease, Amyotrophic Lateral Sclerosis (ALS), Huntington's disease, Parkinson's disease, and many Cancers. The 2000-2006 article by Vijay Pande and Stanford University, http://folding.stanford.edu/, states, "Proteins are biology's workhorses -- its 'nanomachines.' Before proteins can carry out these important functions, they assemble themselves, or 'fold.' The process of protein folding, while critical and fundamental to virtually all of biology, in many ways remains a mystery. Moreover, when proteins do not fold correctly (i.e., 'misfold'), there can be serious consequences, including many well known diseases, such as Alzheimer's, Mad Cow (BSE), CJD, ALS, Huntington's,

Parkinson's disease, and many Cancers and cancer-related syndromes." A copy of the 2000-2006 article by Vijay Pande and Stanford University is enclosed.

Applicants' claimed invention allows the comparison of 3D similarities in protein structure of a first molecule and of a second molecule. When information is known about the first molecule then the information obtained about the comparison with the second molecule provides information about the second molecule. This is "substantial."

In addition, evidence that others have used the invention and have commented favorably on the invention shows that the invention has utility and meets the criteria of being substantial. There are an extensive number of examples of others having used the invention in studying diseases and in developing new pharmaceuticals. Applicants will discuss some examples of these uses; however, the fact that there are extensive uses of the invention demonstrates that that the invention has utility and meets the criteria of being substantial.

The journal article "Consensus sequences improve PSI-BLAST through mimicking profile–profile alignments" by Dariusz Przybylski and Burkhard Rost; in Nucleic Acids Research, 2007, Vol. 35, No. 7, pp. 2238–2246, published online March 16, 2007, under the heading "Evaluation of Alignment Quality" contains the following paragraph:

> "We superposed all models (represented by Ca atom coordinates) with experimentally determined 3D structures using one particular automatic method for structural superposition, namely LGA (40); this method has become one of the standards in the experiments for the Critical Assessment of Structure Prediction [CASP (41)]. First, we computed a Global Distance Test (GDT) (40) that corresponds to the largest, not necessarily continuous subset of residues superimposable within a specified distance threshold. Second, we also computed Longest Continuous Segments (LCS) (40) of residues (consecutively modeled residues) that can fit under a specified RMSD cutoff. The second measure provided us with a local alignment

quality test. Note that we chose a subset of pairs (Q,T) such that for all pairs experimental structures were available; we built the model for Q using the known structure of T and assuming that Q had no known structure, but we evaluated the accuracy of the model using the experimentally known structure for Q. We reported results for two different thresholds. The first was rather stringent (2Å); it focused on the essential core similarities between model and experiment. The second was rather relaxed (5Å) thereby capturing more generic, coarse-grained similarities. Note that GDT computation uses the actual distance threshold while LCS uses average distance (RMSD)."

The fact that the authors Dariusz Przybylski and Burkhard Rost state in the article, ".... one particular automatic method for structural superposition, namely LGA; this method has become one of the standards in the experiments for the Critical Assessment of Structure Prediction" and that the authors have used the invention, is evidence that Applicants' specification which includes the article shows that the invention has utility and meets the criteria of being substantial. A copy of the article, "Consensus sequences improve PSI-BLAST through mimicking profile–profile alignments" by Dariusz Przybylski and Burkhard Rost; in Nucleic Acids Research, 2007, Vol. 35, No. 7, pp. 2238–2246, published online March 16, 2007, is enclosed.

The journal article "Associative memory Hamiltonians for structure prediction without homology: α/β proteins" by Corey Hardin, Michael P. Eastwood, Michael C. Prentiss, Zadia Luthey-Schulten, and Peter G. Wolynes; in Proceedings of the National Academy of Sciences of the United States of America, published online before print January 28, 2003, 10.1073/pnas.252753899, contains the following paragraph:

"The global superposition of two structures can often fail to highlight significant segments of correct native structure. We have submitted the best Q structures to the LGA (Local-Global Alignment) server (http://PredictionCenter.llnl.gov/local/lga), and the results are given in Table 6. The predictions are evaluated according to two measures, LCS and GDT. LCS is the longest continuous (along the sequence) segment that can be

superimposed on the native structure without exceeding a rmsd cutoff. The global distance test (GDT) represents the largest number of residues that lie within a distance cutoff of their correct positions. The set of residues need not be contiguous. In all three cases large portions of the prediction are correct to within the cutoff. We have also used CE to align the predicted and native structure. Note that in all three cases the predicted structure is more similar to the native than any of the database structures, thus demonstrating the ability of the potential to generalize from incorrect scaffolds. The scores of local similarity will, of course, depend on the chosen cutoff. Fig. 2 is a Hubbard plot of the percent of residues below the cutoff, as a function of the cutoff distance."

The fact that the authors Corey Hardin, Michael P. Eastwood, Michael C. Prentiss, Zadia Luthey-Schulten, and Peter G. Wolynes have used the invention, stated in the article as follows: "We have submitted the best Q structures to the LGA (Local-Global Alignment) server (http://PredictionCenter.llnl.gov/local/lga), and the results are given in Table 6," is evidence that the specification which includes the article shows that the invention has utility and meets the criteria of being substantial. A copy of the article, "Associative memory Hamiltonians for structure prediction without homology: α/β proteins" by Corey Hardin, Michael P. Eastwood, Michael C. Prentiss, Zadia Luthey-Schulten, and Peter G. Wolynes; in *Proceedings of the National Academy of Sciences of the United States of America,* published online before print January 28, 2003, 10.1073/pnas.252753899, is enclosed.

The journal article "Crystal Structure of *Mycobacterium tuberculosis* Diaminopimelate Decarboxylase, an Essential Enzyme in Bacterial Lysine Biosynthesis" by Kuppan Gokulan, Bernhard Rupp, Martin S. Pavelka Jr., William R. Jacobs Jr., and James C. Sacchettini; in J. Biol. Chem., Vol. 278, Issue 20, 18588-18596, May 16, 2003, contains the following paragraph:

"Fig. 4. Backbone superposition of known fold type III PLP-dependent enzyme structures. Panel A, color key: cyan, M. tuberculosis DAPDC; magenta, human ODC; green, mouse ODC; and yellow, T. brucei. Panel B, superposition of

M. tuberculosis DAPDC (cyan) with B. stearothermopilus AR (red). The rotation of the AR β-domain relative to the other structures is clearly visible. The superpositions were carried by the Local-Global-Alignment server (Adam Zemla, predictioncenter.llnl.gov/local/lga/lga.html); corresponding r.m.s.d. values are listed in Table III."

The fact that the authors Kuppan Gokulan, Bernhard Rupp, Martin S. Pavelka Jr., William R. Jacobs Jr., and James C. Sacchettini have used the invention, stated in the article as follows: "The superpositions were carried by the Local-Global-Alignment server (Adam Zemla, predictioncenter.llnl.gov/local/lga/lga.html); corresponding r.m.s.d. values are listed in Table III," is evidence that the specification which includes the article shows that the invention has utility and meets the criteria of being substantial. A copy of the article, "Crystal Structure of *Mycobacterium tuberculosis* Diaminopimelate Decarboxylase, an Essential Enzyme in Bacterial Lysine Biosynthesis" by Kuppan Gokulan, Bernhard Rupp, Martin S. Pavelka Jr., William R. Jacobs Jr., and James C. Sacchettini; in J. Biol. Chem., Vol. 278, Issue 20, 18588-18596, May 16, 2003, is enclosed.

The journal article "Structural and Functional Basis of CXCL12 (Stromal Cell-derived Factor-1α) Binding to Heparin" by James W. Murphy, Yoonsang Cho, Aristidis Sachpatzidis, Chengpeng Fan, Michael E. Hodsdon, and Elias Lolis; in THE JOURNAL OF BIOLOGICAL CHEMISTRY VOL. 282, NO. 13, pp. 10018–10027, March 30, 2007, contains the following paragraph:

"The crystal structure of the complex is shown in Fig. 5A. As in the native structure three _-strands from each subunit form an extended six-stranded _-sheet that in turn forms a pocket where a heparin disaccharide molecule is located. Root mean square deviation values were determined for each _ carbon between the apo and bound structures using a local global alignment similarity calculation (44). 44. Zemla, A. (2003) Nucleic Acids Res. 31, 3370–3374."

The fact that the authors James W. Murphy, Yoonsang Cho, Aristidis Sachpatzidis, Chengpeng Fan, Michael E. Hodsdon, and Elias Lolis have used

the invention, stated in the article as follows: in the article, "....using a local global alignment similarity calculation (44). 44. Zemla, A. (2003) Nucleic Acids Res. 31, 3370–3374,"is evidence that Applicants' specification which includes the article shows that the invention has utility and meets the criteria of being substantial. A copy of the article, "Structural and Functional Basis of CXCL12 (Stromal Cell-derived Factor-1α) Binding to Heparin" by James W. Murphy, Yoonsang Cho, Aristidis Sachpatzidis, Chengpeng Fan, Michael E. Hodsdon, and Elias Lolis; in THE JOURNAL OF BIOLOGICAL CHEMISTRY VOL. 282, NO. 13, pp. 10018–10027, March 30, 2007, is enclosed.

The journal article "On the Accuracy of Homology Modeling and Sequence Alignment Methods Applied to Membrane Proteins" by Lucy R. Forrest, Christopher L. Tang, and Barry Honig; in Biophysical Journal, Volume 91, July 2006, pp. 508–517, contains the following paragraph:

> "Gap penalties were also assigned according to the location of core regions or secondary-structure elements. We used the local-global alignment method where unaligned terminal residues are only penalized in the query. 39. Zemla, A. 2003. LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Res. 31:3370–3374."

The fact that the authors Lucy R. Forrest, Christopher L. Tang, and Barry Honi have used the invention, stated in the article as follows: "We used the local-global alignment method ...," is evidence that Applicants' specification which includes the article shows that the invention has utility and meets the criteria of being substantial. A copy of the article, "On the Accuracy of Homology Modeling and Sequence Alignment Methods Applied to Membrane Proteins" by Lucy R. Forrest, Christopher L. Tang, and Barry Honig; in Biophysical Journal, Volume 91, July 2006, pp. 508–517, is enclosed.

The journal article "Antibody Elbow Angles are Influenced by their Light Chain Class" by Robyn L. Stanfield, Adam Zemla, Ian A. Wilson, and Bernhard Rupp; in J. Mol. Biol. (2006) 357, 1566–1574, contains the following paragraph:

> "The program RBOW is implemented on a Linux platform (Apache web server) using a combination of Perl scripts and standard Fortran90 code. The alignment program used is the local-global alignment program LGA, which assures minimal dependence of the results on the Fab numbering convention used. The program allows upload of PDB format coordinate files, selection of heavy and light chain identifiers, and input of domain boundaries, for which reasonable default values are provided."

The fact that the authors Robyn L. Stanfield, Adam Zemla, Ian A. Wilson, and Bernhard Rupp have used the invention, stated in the article as follows: "The alignment program used is the local-global alignment program LGA," is evidence that Applicants' specification which includes the article shows that the invention has utility and meets the criteria of being substantial. A copy of the article, "Antibody Elbow Angles are Influenced by their Light Chain Class" by Robyn L. Stanfield, Adam Zemla, Ian A. Wilson, and Bernhard Rupp; in J. Mol. Biol. (2006) 357, 1566–1574, is enclosed.

The journal article "Mycobacterium tuberculosis RmlC epimerase (Rv3465): a promising drug-target structure in the rhamnose pathway" by Katherine A. Kantardjieff, Chang-Yub Kim, Cleo Naranjo, Geoffry S.Waldo, Timothy Lekin, Brent W. Segelke, Adam Zemla, Min S. Park, Thomas C. Terwilliger, and Bernhard Rupp; in Acta Cryst. (2004). D60, 895–902, contains the following paragraph:

> "Figure 1 - Pairwise structural alignment of homologous protein chains with RmlC from MTB using the local global alignment program LGA (Zemla, 2003)."

The fact that the authors Katherine A. Kantardjieff, Chang-Yub Kim, Cleo Naranjo, Geoffry S.Waldo, Timothy Lekin, Brent W. Segelke, Adam Zemla, Min S. Park, Thomas C. Terwilliger, and Bernhard Rupp have used the invention, stated in the article as follows: ".... using the local global alignment program LGA (Zemla, 2003)," is evidence that Applicants' specification which includes the article shows that the invention has utility and meets the criteria of being substantial. A copy of the article, "Mycobacterium tuberculosis RmlC epimerase (Rv3465): a promising drug-target structure in the rhamnose pathway" by Katherine A. Kantardjieff, Chang-Yub Kim, Cleo Naranjo, Geoffry S.Waldo, Timothy Lekin, Brent W. Segelke, Adam Zemla, Min S. Park, Thomas C. Terwilliger, and Bernhard Rupp; in Acta Cryst. (2004). D60, 895–902 is enclosed.

35 USC § 102(b) Rejection

In the Office Action mailed January 4, 2007, claims 1, 2, 4-7, and 10-13 were rejected under 35 U.S.C. § 102(b) as being anticipated by the Lackner et al reference (ProSup: a refined tool for protein structure alignment by Peter Lackner, Walter A. Koppensteiner, Manfred J. Sippl, and Francisco S. Domingues in Protein Eng. 2000 13: 745-752; doi:10.1093/protein/13.11.745).

The standard for a 35 U.S.C. §102 rejection is stated in RCA Corp. v. Applied Digital Systems, Inc, 221PQ 385, 388 (d. Cir. 1984) "Anticipation is established only when a single prior art reference discloses, either expressly or under principles of inherency, each and every element of a claimed invention."

Applicants submit that the invention claimed in claims 1, 2, 4-7, and 10-13 is not anticipated by the Lackner et al reference. Applicants point out that the following elements of Applicants' claims 1, 2, 4-7, and 10-13 are not found in the Lackner et al reference:

"independently comparing the first molecule and the second molecule using said preselected information and using Longest Continuous Segments (LCS) analysis," or

"independently comparing the first molecule and the second molecule using said preselected information and using Global Distance Test (GDT) analysis," or

"independently comparing the first molecule and the second molecule using said preselected information and using Local Global Alignment Scoring function (LGA_S) analysis to provide a calculated alignment between said first molecule and said second molecule," or

"repeating said steps to find all the regions of 3D similarities between considered protein structures," or

"generating an output containing information about said calculated alignment," or

"comparing the first molecule and the second molecule using said preselected information and Longest Continuous Segments (LCS) analysis," or

"comparing the first molecule and the second molecule using said preselected information and Global Distance Test (GDT) analysis," or

"comparing the first molecule and the second molecule using said preselected information and Local Global Alignment Scoring function (LGA_S) analysis," or

"repeating said steps to find all the regions of 3D similarities between considered protein structures," or

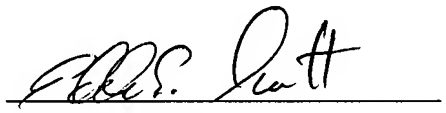
"generating the output from the program containing the complete information about the quality of the calculated alignment which includes distances between the corresponding residues, LCS data and GDT data."

Since the elements described above are not found in the Lackner et al reference, the Lackner et al reference does not support a 35 U.S.C. §102(e) rejection of Applicants' claims 1, 2, 4-7, and 10-13 and the rejection should be withdrawn.

SUMMARY

The undersigned respectfully submits that, in view of the foregoing amendments and the foregoing remarks, the rejections of the claims raised in the Office Action dated January 4, 2007 have been fully addressed and overcome, and the present application is believed to be in condition for allowance. It is respectfully requested that this application be reconsidered, that the claims be allowed, and that this case be passed to issue. If it is believed that a telephone conversation would expedite the prosecution of the present application, or clarify matters with regard to its allowance, the Examiner is invited to call the undersigned attorney at (925) 424-6897.

Respectfully submitted,



Eddie E. Scott
Attorney for Applicant
Registration No. 25,220
Tel. No. (925) 424-6897

Livermore, California
Dated: June 21, 2007

3370–3374 Nucleic Acids Research, 2003, Vol. 31, No. 13
DOI: 10.1093/nar/gkg571

LGA: a method for finding 3D similarities in protein structures

Adam Zemla*

Lawrence Livermore National Laboratory, 7000 East Avenue, Livermore, CA 94550, USA

Received February 16, 2003; Revised and Accepted April 4, 2003

ABSTRACT

We present the LGA (Local-Global Alignment) method, designed to facilitate the comparison of protein structures or fragments of protein structures in sequence dependent and sequence independent modes. The LGA structure alignment program is available as an online service at http://PredictionCenter.llnl.gov/local/lga. Data generated by LGA can be successfully used in a scoring function to rank the level of similarity between two structures and to allow structure classification when many proteins are being analyzed. LGA also allows the clustering of similar fragments of protein structures.

INTRODUCTION

If one were to compare two slightly different conformations of the same protein, the overall root mean square deviation (RMSD) of all corresponding C-alpha atoms would give a useful impression of the similarity between the two structures. Unfortunately, a small perturbation in just one part of the protein (e.g. in a hinge joining two domains) can create a large RMSD and it would seem that the two structures are very different overall. Thus, it is desirable to also consider local regions of the proteins in assessing their similarity. In essence, the smaller such 'deviant' regions, the more similar the two structures are. If one compares two different proteins, where there is not a preassigned correspondence between amino acid residues, a sequence-independent alignment (residue correspondence) has to be generated first, adding another significant level of complexity.

We were thus motivated to develop a method that would take into account both local and global structure superpositions and also would be capable of working without a preassigned residue correspondence. We called this method 'LGA' for local/global alignment. Below we describe our algorithm and apply the LGA program to several test cases in order to highlight some of its features.

EVALUATING STRUCTURE SIMILARITY BETWEEN PROTEINS

Most structure comparison programs are built on the principle that a suitable scoring function can be defined with its optimum corresponding to the most significant structural match for a given protein. Many established comparison techniques evaluate structural similarity by two numbers, the RMSD between two superimposed structures together with the number of 'equivalent' (structurally aligned) residues. However, it is very difficult to optimize these two quantities simultaneously, since one can be optimized at the expense of the other. For example, the structural aligner, DALI (1), which is based on the alignment of distance matrices, solves the optimization problem by combining several numbers into a single quantity, called z-score. ProSup (2) maximizes the number of equivalent residues while RMSD is kept close to a constant value. An additional problem can arise when structures are similar in small, local regions. These regions of similarity can be overlooked when one global superposition is applied. In general, in many cases there is no 'best' superposition that reveals all regions of similarity between compared proteins.

To resolve these problems while comparing two structures, the LGA program generates many different local superpositions to detect regions where proteins are similar. The LGA scoring function has two components, LCS (longest continuous segments) and GDT (global distance test), established for the detection of regions of local and global structure similarities between proteins. These two measures were extensively tested during the last three successive rounds of CASP [Critical Assessment of Techniques for Protein Structure Prediction (3–7)] providing constructive ranking of evaluated 3D models. In comparing two protein structures, the LCS procedure is able to localize and superimpose the longest segments of residues that can fit under a selected RMSD cutoff. The GDT algorithm is designed to complement evaluations made with LCS searching for the largest (not necessary continuous) set of 'equivalent' residues that deviate by no more than a specified *distance* cutoff.

Data generated by the LCS and GDT algorithms

In an attempt to generate detailed information about regions of local similarity between two protein structures (Molecule1 and Molecule2) or segments thereof, each residue from Molecule2 is assigned to the largest set of residue pairs (C-alpha atoms from Molecule1 and Molecule2) provided it is a part of that set and can be fit under a selected RMSD (LCS algorithm) or distance (GDT algorithm) cutoff. If an analysis of two structures is based only on the superpositions limited to one

*Tel: +1 925 423 5571; Fax: +1 925 422 2133; Email: adamz@llnl.gov

Published by Oxford University Press 2003

Table 1. Example of data generated by LCS and GDT analyses

Column #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	...
Cutoffs:					1 Å	2 Å	5 Å	0.5 Å	1.0 Å	1.5 Å	2.0 Å	2.5 Å	3.0 Å	3.5 Å	4.0 Å	4.5 Å	...
LCS_GDT	Molecule-1		Molecule-2		Length_of_the												
LCS_GDT	Residue		Residue		continuous												
LCS_GDT	Name	Number	Name	Number	Segment			Global distance test data									
LCS_GDT	V	40	A	29	23	26	90	10	18	22	23	24	24	27	33	49	...
LCS_GDT	A	41	Q	30	23	26	90	10	18	22	23	24	25	27	42	55	...
LCS_GDT	L	42	L	31	23	26	90	4	7	20	23	24	25	36	46	55	...
LCS_GDT	E	43	E	32	8	26	90	4	7	15	23	24	25	35	46	55	...
LCS_GDT	Q	44	V	33	8	26	90	4	6	9	18	24	26	37	46	55	...
LCS_GDT	T	45	T	34	8	26	90	4	7	9	13	22	25	36	46	55	...
LCS_GDT	G	46	G	35	8	14	90	3	7	9	12	17	22	35	46	55	...

selected RMSD or distance cutoff then it would not give full information on similarity between the two structures; some similarities would be detected, some would not. To avoid such limitations, LCS results are generated for a set of increasing RMSD cutoffs [1 Å (Ångstrom), 2 Å and 5 Å], and in the GDT analysis, two structures are scanned every 0.5 Å, starting from 0.5 Å up to a 10.0 Å distance cutoff. This approach allows us to gather very detailed information on local similarities between two structures. The results of such calculations are reported in the format as shown in Table 1.

In the output shown in Table 1, columns 2–5 provide information on residues from two compared structures, and columns 6, 7 and 8 show the results from LCS analyses under 1 Å, 2 Å and 5 Å RMSD cutoffs, respectively. For example, residue L-31 from Molecule2 is a member of a 23-residue long continuous segment that can be superimposed with corresponding residues from Molecule1 under a 1 Å RMSD cutoff, but residue E-32 is an element of a segment consisting of just eight residues at an RMSD cutoff of 1 Å. In columns 9–28 the results of GDT analysis under 0.5 Å through 10.0 Å distance cutoffs are reported. For example, residue E-32 belongs to a set of four residues (not necessarily continuous) that can fit under a 0.5 Å distance cutoff, a set of seven residues under a 1.0 Å and a 25-residue set under 3.0 Å.

The GDT algorithm

In the GDT procedure, the search for an optimal superposition between two structures is performed as follows. For each selected pair of three, five and seven residue-long segments from both structures, an RMSD and a superposition are calculated. Each calculated superposition is used as a starting point to give an initial list of equivalent residues (C-alpha atom pairs from Molecule1 and Molecule2). The list of such equivalences is iteratively extended to collect the largest set of residues that can fit under a given distance cutoff. The goal of the iterative procedure is to exclude atoms that are more distant than a threshold (distance cutoff) between Molecule1 and Molecule2 after the transform is applied. Starting from the initial set of atom pairs, the algorithm is as follows: (a) obtain the transform; (b) apply the transform; (c) identify all atom pairs for which the distance is larger than the threshold; (d) re-obtain the transform, excluding those atoms; (e) repeat steps (b)–(d) until the set of atoms used in calculations is the same for two cycles running.

The LCS and GDT algorithms are complementary

Results of the LCS algorithm identify local regions of similarity between proteins, while residues identified by GDT arise from anywhere in the structure (i.e. sequence continuity need not be maintained). From this point of view, GDT detects global, as opposed to local, similarity. Using GDT we focus on distance rather than RMSD. Using LCS, however, we can optimize (minimize) RMSD on the selected residues. So from this point of view, LCS gives complete and optimal information. Working with distance analysis (maximum norm) an optimal method for finding the 'best superposition', which will minimize the distances between all selected residues, is not known. Results can only be approximated. So to find the 'best' global structural match, GDT uses many distance cutoffs and superpositions. The GDT algorithm 'tests' each residue one by one from Molecule2, trying to assign it to the largest set of residues possible (not necessarily continuous) deviating from Molecule1 by no more than a specified distance cutoff. GDT evaluates a selected but large number of superpositions, in effect yielding consistently reliable results.

Description of the LGA scoring function

By combining these two techniques (RMSD based and distance based), LGA not only calculates a 'best' superposition between two proteins (meaning 'under certain RMSD and distance cutoffs'), but also identifies the regions of local similarity between compared structures. In the structure alignment search procedure, for each generated list of equivalent residues, the following values are calculated: LCS_*vi*—percent of residues (continuous set) that can fit under an RMSD cutoff of *vi* Å (for $vi = 1.0, 2.0, \ldots$) and GDT_*vi*—an estimation of the percent of residues (largest set) that can fit under the distance cutoff of *vi* Å (for $vi = 0.5, 1.0, \ldots$). A scoring function (LGA_S) can be defined as a combination of these values and can be used to evaluate the level of structure similarity of selected regions. For a given parameter w ($0.0 \leq w \leq 1.0$), representing a weighting factor, we calculate LGA_S by the

Table 2. NMR models 1m2f_A_1–1m2f_A_25 compared to an average model 1m2e_A and sorted by GDT_TS value where GDT_TS = (P1 + P2 + P4 + P8)/4, and Pd is a percent of residues from 1m2e_A that can be superimposed with corresponding residues from 1m2f_A_n under selected distance cutoffs d = 1, 2, 4, 8

Model	N1	N2	DIST	N	RMSD	GDT_TS
1m2f_A_8	135	135	3.0	135	0.79	97.037
1m2f_A_16	135	135	3.0	133	0.70	96.296
1m2f_A_17	135	135	3.0	133	0.80	96.296
1m2f_A_2	135	135	3.0	135	0.91	96.296
1m2f_A_1	135	135	3.0	133	0.93	96.111
1m2f_A_19	135	135	3.0	134	0.95	96.111
1m2f_A_11	135	135	3.0	134	0.84	95.926
1m2f_A_14	135	135	3.0	133	0.91	95.926
1m2f_A_20	135	135	3.0	133	0.94	95.926
1m2f_A_7	135	135	3.0	131	0.85	95.741
1m2f_A_21	135	135	3.0	130	0.80	95.556
1m2f_A_5	135	135	3.0	134	1.04	95.556
1m2f_A_10	135	135	3.0	135	1.09	95.556
1m2f_A_18	135	135	3.0	134	0.89	95.370
1m2f_A_12	135	135	3.0	133	0.92	95.370
1m2f_A_13	135	135	3.0	131	0.95	95.370
1m2f_A_15	135	135	3.0	130	0.80	95.185
1m2f_A_24	135	135	3.0	133	0.89	95.185
1m2f_A_22	135	135	3.0	131	0.85	95.000
1m2f_A_25	135	135	3.0	134	0.94	95.000
1m2f_A_9	135	135	3.0	132	1.14	95.000
1m2f_A_4	135	135	3.0	130	1.01	94.444
1m2f_A_3	135	135	3.0	129	0.74	94.074
1m2f_A_23	135	135	3.0	132	1.00	93.704
1m2f_A_6	135	135	3.0	130	1.05	92.963

formula: $LGA\_S = w * S(GDT) + (1 - w) * S(LCS)$ where S(F) function is defined as follows:

```
foreach vi (v1, v2, ..., vk) {
    Y=(k-i+1)/k; X=X+Y*F_vi;
}
S(F=X/((1+k)*k/2);
```

The same scoring function is applied by the LGA program to perform the selection and ranking of the regions of structure similarities in the sequence dependent mode of analysis as well as in the sequence independent mode.

Graphical presentation of results from structure comparison of NMR models

How can the results of a multiple superposition (Table 1) between two structures be visualized? Let us compare an NMR average model, 1m2e_A, of the N-terminal domain of *Synechococcus elongatus kaia* (KAIA135N) with its 25-member family of low energy (designated 1m2f_A_n). In Table 2, NMR models are sorted by GDT_TS values.

In Figure 1 we show how colored strip charts can be used to plot output from the LGA program (data from Tables 1 and 2). Each bar from Figure 1A or B corresponds to one pair of analyzed structures. The ordering of bars is the same as in Table 2. Rasmol plots (Fig. 1C and D) are provided only for one model, 1m2f_A_2 (fourth in Table 2 and bar charts).

Figure 1B shows that the results of multi-superposition LGA analysis as reported in Table 1 can be used to detect regions of similarity between proteins from those where the structures differ. Analysis based on a single superposition (Fig. 1A) does not distinguish the regions of similarity so clearly.

Table 3. List of the 10 of the closest PDB structures to 1m2e_A found by the LGA program. Proteins are sorted by N—the number of superimposed residues under a distance cutoff 5.0 Å

Name	N1	N2	DIST	N	RMSD	Seq_Id	LGA_S
1a04_B	205	135	5.0	118	2.36	11.86	63.707
1a2o_B	347	135	5.0	117	2.47	11.97	62.598
1rnl	200	135	5.0	116	2.14	12.07	69.416
1e6m_A	128	135	5.0	116	2.23	10.34	64.587
6chy_A	128	135	5.0	116	2.25	10.34	63.363
6chy_B	128	135	5.0	116	2.26	10.34	64.196
2che	128	135	5.0	116	2.28	9.48	64.372
1a0o_C	128	135	5.0	116	2.29	10.34	63.826
1ffg_C	128	135	5.0	116	2.29	10.34	63.161
1ffw_A	128	135	5.0	116	2.32	9.48	62.522

Graphical presentation of results from sequence independent database searches

The greatest utility of structure alignment programs, such as LGA, lies in their ability to superimpose protein structures regardless of sequence identity and to detect regions of structural similarity. In Table 3 we provide a list of 10 of the closest PDB structural matches to the already mentioned NMR average model 1m2e_A (CASP5 target T0138). The PDB database search was performed with the use of the LGA program working in sequence independent mode. The level of sequence identity (Seq_Id) to other structurally similar PDB entries was very low, on the order of 12%.

Graphical presentation of the results from the LGA database search is given in Figure 2. Each bar corresponds to one hit to a protein from the PDB database. The bars are ordered as in Table 3. Figure 2A shows regions of structural similarity (in green) between the reference structure, 1m2e_A and each PDB database hit from Table 3. Regions of high structural diversity are shown in red. A RasMol plot (Fig. 2B) is given for the best database match, PDB protein, 1a04_B.

LGA IN COMPARISON WITH OTHER PROGRAMS

An important requirement for any structure comparison method is its ability to detect weak structural similarity. In the Table 4 we compare results of LGA to those of four methods available as web services and which are frequently used by the scientific community: VAST (8), DALI (1), CE (9) and ProSup (10). This identical dataset was used in a comparison of ProSup to other structural alignment programs [Table III in reference (10)].

The number N of structurally equivalent residues differs considerably for several protein pairs. One would expect that a higher number of equivalent residues would indicate better performance of a particular method in the detection of structural similarity. However, comparing the number of equivalent residues is insufficient without taking RMSD into account. RMSD reported by LGA is fairly constant in all cases. Our program can keep the smallest range of RMSD 1.9–2.6 while providing a high number of aligned residues. In a

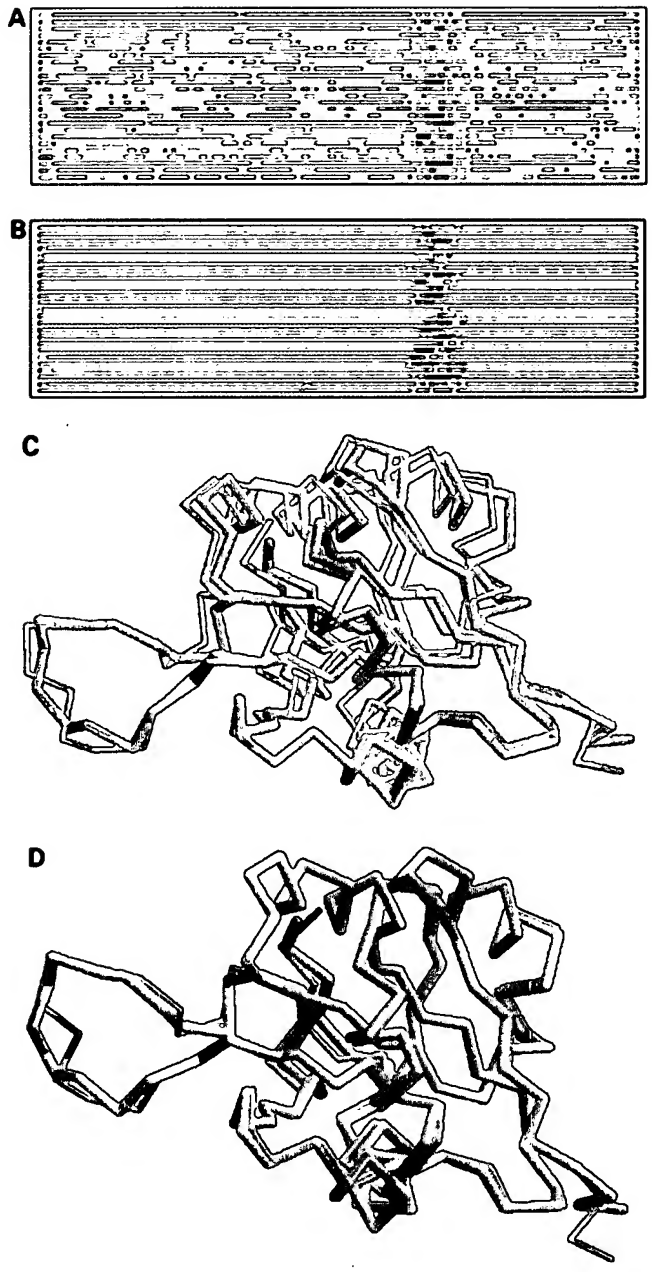


Figure 1. (**A**) C-alpha–C-alpha distance deviation bars from one LGA superposition under a 3.0 Å distance cutoff. Residues superimposed below 1.0 Å are in green, below 2.0 Å in yellow, below 3.0 Å in orange, below 4.0 Å in brown and residues at or above 4.0 Å in red. (**C**) RasMol plot of two superimposed structures: 1m2f_A_2 and 1m2e_A. Colors correspond to the fourth bar from (A). (**B**) C-alpha–C-alpha deviation bars for multiple LGA superpositions. (**D**) RasMol plot of superimposed structures 1m2f_A_2 and 1m2e_A corresponding to fourth bar representation from (B) where >85.0% of equivalent residues under distance cutoff = 1.5 Å are in green, >70.0%: yellow, >50.0%: orange, >20.0%: brown and ≤20.0%: red.

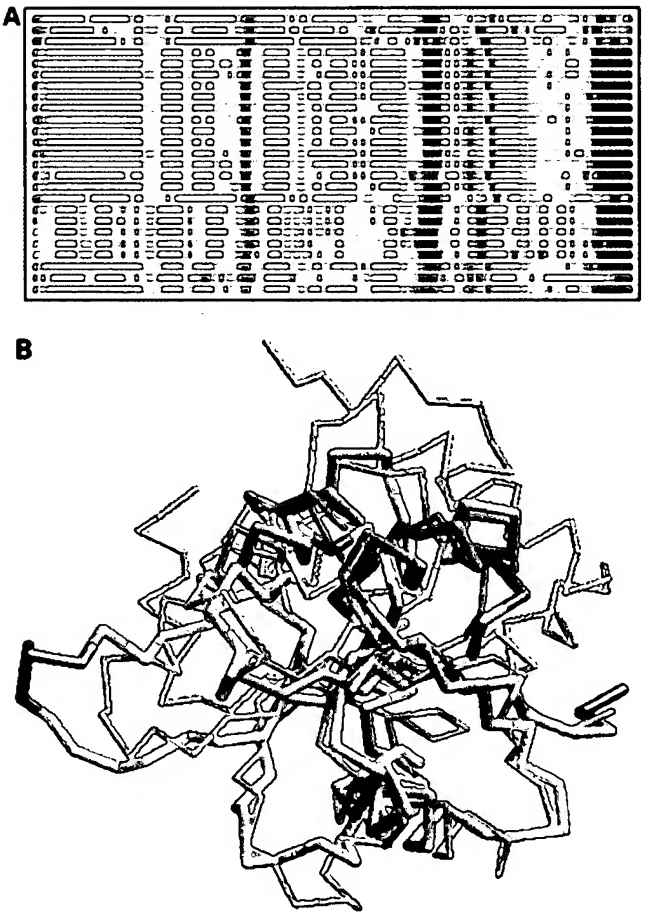


Figure 2. Bar representation of the results from sequence independent LGA superpositions, and a RasMol plot of superimposed first template 1a04_B and T0138. Residues superimposed below 2.0 Å are in green, below 4.0 Å in yellow, below 6.0 Å in orange and residues at or above 6.0 Å or not superimposed are in red (target) and in white (template).

Table 4. Comparison of structure alignments for 10 'difficult' structures (11). For each protein pair the N and RMSD results from different methods are provided where N is a number of equivalent residues with the corresponding RMSD

Proteins		VAST	DALI	CE	ProSup	LGA
1bge-B	2gmf-A	71/2.3	94/3.3	107/3.9	87/2.4	91/2.5
1cew-I	1mol-A	75/2.0	81/2.3	81/2.3	76/1.9	79/2.0
1cid	2rhe	78/2.0	96/3.1	97/2.9	84/2.3	93/2.3
1crl	1ede	186/3.7	212/3.6	219/3.8	161/2.6	182/2.6
1fxi-A	1ubq	48/2.1	52/2.5	64/3.8	54/2.6	61/2.6
1ten	3hhr-B	76/1.5	86/1.9	87/1.9	85/1.7	87/1.9
1tie	4fgf	76/1.6	114/3.1	116/2.9	101/2.4	104/2.3
2sim	1nsb-A	299/4.2	289/3.2	275/3.0	248/2.6	269/2.6
2aza-A	1paz	70/2.1	82/3.0	84/2.9	82/2.6	80/2.2
3hla-B	2rhe	58/2.3	74/3.0	83/3.3	71/2.7	74/2.5

comparison to ProSup, in some cases LGA superimposes more residues under the same distance cutoff (sometimes with a slightly higher value of RMSD). During the CASP4 competition, both programs were used for evaluation of structure predictions and to perform PDB searches showing similar results.

CONCLUSION

Optimizing the number of equivalent residues while keeping the RMSD constant provides a simple and intuitive measure of structure similarity (as concluded in 10). Such a measure can be used effectively for ranking in database searches. We show that in LGA an additional requirement of fulfilling distance restrictions combined with extensive analysis of regions of local similarities (from searches with multiple distance and RMSD cutoffs) was successfully implemented. Our approach can generate data that provide detailed information not only about the degree of global similarity but also about regions of local similarity in protein structures. It allows the clustering of similar fragments of structures, and the use of such clusters to identify sequence patterns that would represent local structural motifs.

Accessibility, limitations and further development of the program

An online LGA service is accessible at http://PredictionCenter.llnl.gov/local/lga. The required input consists of two sets of protein structure coordinates in PDB format. For calculations, a user can specify chains, residue segments or select isolated residues. As a result of LGA processing the user will get the translation/rotation matrices, the rotated coordinates of the first structure and (optionally) the coordinates of the second structure (target, unchanged). Depending on need, the user can choose between several options described in detail in the 'help' file. For example, there are four options: -1, -2, -3, -4 that allow the user to select the calculation method. Option-1 is a standard RMSD calculation performed on all selected residues in both structures. Option-2 allows the selection of a user specified distance cutoff (-d:f.f), and only the residues within this distance cutoff will be superimposed using an iterative procedure as described in the section 'The GDT algorithm'. Option-3 is used to generate detailed LCS and GDT information about regions of local and global similarity as shown in Table 1 (see section 'Data generated by the LCS and GDT algorithms'). And finally, option-4 is used to perform the structure alignment search (structure comparison of proteins without a preassigned residue correspondence). With option '-d:f.f', which specifies a distance cutoff in Ångstroms, the user may force LGA to calculate tighter or more relaxed superpositions for a selected region. The possible ranges for distance cutoff are from 0.1 to 10.0 Å. The default value is 5 Å. For a description of more advanced options please consult the online documentation.

The program reports a single, final superposition and no alternative alignments are provided. In the current version of the LGA server, a text-only output is available. A future release of the service will contain a graphical presentation package to generate plots as shown in Figures 1 and 2.

ACKNOWLEDGEMENTS

The author thanks John Moult for establishing the CASP experiment during which the LGA program was extensively tested. The CASP assessors have contributed valuable comments on the use of the results from LGA for CASP evaluations. The author also wishes to thank Michael Levitt for his valuable suggestions on the LGA program and its performance. Finally, the author thanks his colleagues from LLNL: Tom Slezak for his continuous help and support, Ceslovas Venclovas and Daniel Barsky for their comments in using the program and Carol Zhou and Dorota Sawicka for critical reading of the manuscript. This work was performed under the auspices of the US Department of Energy by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48.

REFERENCES

1. Holm,L. and Sander,C. (1993) Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, **233**, 123–138.
2. Feng,Z.K. and Sippl,M.J. (1996) Optimum superimposition of protein structures: ambiguities and implications. *Fold Des.*, **1**, 123–132.
3. Zemla,A., Venclovas,C., Reinhardt,A., Fidelis,K. and Hubbard,T.J. (1997) Numerical criteria for the evaluation of ab initio predictions of protein structure. *Proteins*, **S1**, 140–150.
4. Zemla,A., Venclovas,C., Moult,J. and Fidelis,K. (1999) Processing and analysis of CASP3 protein structure predictions. *Proteins*, **S3**, 22–29.
5. Orengo,C.A., Bray,J.E., Hubbard,T., LoConte,L. and Sillitoe,I. (1999) Analysis and assessment of *ab initio* three-dimensional prediction, secondary structure, and contacts prediction. *Proteins*, **S3**, 149–170.
6. Zemla,A., Venclovas,C., Moult,J. and Fidelis,K. (2001) Processing and evaluation of predictions in CASP4. *Proteins*, **45** (Suppl. 5), 13–21.
7. Tramontano,A., Leplae,R. and Morea,V. (2001) Analysis and assessment of comparative modeling predictions in CASP4. *Proteins*, **45** (Suppl. 5), 22–38.
8. Gibrat,J-F., Madej,T. and Bryant,S.H. (1996) Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.*, **6**, 377–385.
9. Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, **11**, 739–747.
10. Lackner,P., Koppensteiner,W.A., Sippl,M.J. and Domingues,F.S. (2000) ProSup: a refined tool for protein structure alignment. *Protein Eng.*, **13**, 745–752.
11. Fischer,D., Eloffson,A., Rice,D.W. and Eisenberg,D. (1996) *Proceedings of the 1st Pacific Symposium on Biocomputing*. In Hunter,L. and Klein,T. (eds), World Scientific Publishing Company, Singapore, pp. 300–318.

Our goal: to understand protein folding, misfolding, and related diseases

What is protein folding and how is folding linked to disease? Proteins are biology's workhorses -- its "nanomachines." Before proteins can carry out these important functions, they assemble themselves, or "fold." The process of protein folding, while critical and fundamental to virtually all of biology, in many ways remains a mystery.

Moreover, when proteins do not fold correctly (i.e. "misfold"), there can be serious consequences, including many well known diseases, such as Alzheimer's, Mad Cow (BSE), CJD, ALS, Huntington's, Parkinson's disease, and many Cancers and cancer-related syndromes.

You can help by simply running a piece of software. Folding@Home is a distributed computing project -- people from through out the world **download** and run software to band together to make one of the largest supercomputers in the world. Every computer makes the project closer to our goals.

Folding@Home uses novel computational methods coupled to distributed computing, to simulate problems thousands to millions of times more challenging than previously achieved.

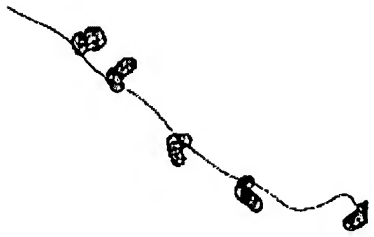
What have we done so far? We have had several successes. You can read about them on our Science page, Results section, or go directly to our press and papers page. **Want to learn more?** Click on the links on the left for downloads or more information. You can also download our Executive Summary, which is a PDF suitable for distribution. Also, you can learn more by watching recent seminars (Stanford BMI ; Xerox PARC). One can also help by donating funds to the project, via Stanford University.

NEW! **Folding@Home *PS3* client now available.**

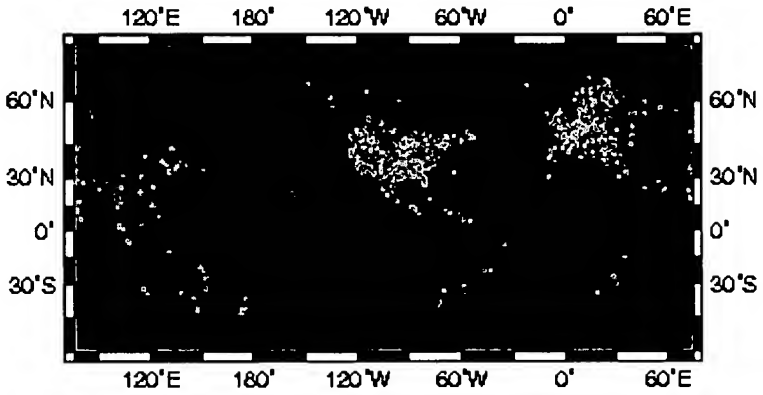
CURRENT PROJECTS AND PROGRESS TO DATE:

- **Alzheimer's Disease**
- **Cancer**
- **Huntington's Disease**
- **Osteogenesis Imperfecta**
- **Parkinson's Disease**
- **Ribosome & antibiotics**

Your participation can help lead to a cure for these diseases.
Click here for free download.



Results from Folding@Home



Since October 1, 2000, over 1,000,000 CPUs throughout the world have participated in Folding@Home. Each additional CPU gives us an added boost in performance, allowing us to tackle more difficult problems or solve existing research faster or more accurately.

http://folding.stanford.edu/

2238–2246 Nucleic Acids Research, 2007, Vol. 35, No. 7 Published online 16 March 2007
doi:10.1093/nar/gkm107

Consensus sequences improve PSI-BLAST through mimicking profile–profile alignments

Dariusz Przybylski[1,2,*] and Burkhard Rost[1,2,3]

[1]Department of Biochemistry and Molecular Biophysics, Columbia University, 630 West 168th Street, New York, NY 10032, USA, [2]Columbia University Center for Computational Biology and Bioinformatics (C2B2), 1130 St. Nicholas Ave. Rm. 801, New York, NY 10032, USA and [3]NorthEast Structural Genomics Consortium (NESG), Columbia University, 1130 St. Nicholas Ave. Rm. 802, New York, NY 10032, USA

Received December 20, 2006; Revised February 5, 2007; Accepted February 6, 2007

ABSTRACT

Sequence alignments may be the most fundamental computational resource for molecular biology. The best methods that identify sequence relatedness through profile–profile comparisons are much slower and more complex than sequence–sequence and sequence–profile comparisons such as, respectively, BLAST and PSI-BLAST. Families of related genes and gene products (proteins) can be represented by consensus sequences that list the nucleic/amino acid most frequent at each sequence position in that family. Here, we propose a novel approach for consensus-sequence-based comparisons. This approach improved searches and alignments as a standard add-on to PSI-BLAST without any changes of code. Improvements were particularly significant for more difficult tasks such as the identification of distant structural relations between proteins and their corresponding alignments. Despite the fact that the improvements were higher for more divergent relations, they were consistent even at high accuracy/low error rates for non-trivially related proteins. The improvements were very easy to achieve; no parameter used by PSI-BLAST was altered and no single line of code changed. Furthermore, the consensus sequence add-on required relatively little additional CPU time. We discuss how advanced users of PSI-BLAST can immediately benefit from using consensus sequences on their local computers. We have also made the method available through the Internet (http://www.rostlab.org/services/consensus/).

INTRODUCTION

Improved database search and alignment methods boost biology

Sequence alignments are fundamental to modern molecular biology. They are used to detect evolutionary relationships among proteins and genes; they also provide the basis for most advanced predictions of structure and function for biomolecules. The more organisms are sequenced, the more the need for sensitive and accurate database search and alignment methods increases. In conjunction with an appropriate scoring (decision) function, sequence alignment methods can often distinguish homologous from non-homologous genes/proteins. Alignments are also used to establish residues that are conserved between related sequences. This helps to identify residues that are most important for function and to transfer three-dimensional (3D) coordinates in comparative modeling of protein structures. Since most relations between genes or proteins are observed at large evolutionary distances, small improvements in the sensitivity and accuracy of database searches and alignments may translate to thousands of novel annotations that could guide and accelerate experimental biology.

PSI-BLAST strikes a very good compromise between speed and sensitivity

Ideally, an alignment method should accurately identify and align related sequences in today's rapidly expanding databases within the shortest possible time. While we want to simultaneously optimize speed and reliability, in practice, there is a tradeoff; more accurate alignment methods are relatively slow (e.g. profile–profile alignment algorithms), while very fast methods are far less sensitive than we might wish [e.g. BLAST (1)]. Generally, the most

*To whom correspondence should be addressed. Tel: +1 212 851 4669; Fax: +1 212 851 5176; Email: dsp23@columbia.edu

sensitive and accurate methods use profile–profile comparisons (2–5). In those algorithms, nucleic/amino acid substitution patterns are used for both sequences being aligned. One downside of profile–profile alignments is that they are relatively slow. When aligning two sequences of lengths m and n they require on the order of S^*m^*n operations (where S is the size of sequence alphabet—20 for proteins). Moreover, the algorithm is not easily amenable to acceleration. In contrast, the less powerful sequence–profile alignment methods can be easily accelerated. This is most impressively visible in the case of PSI-BLAST (6) that combines techniques for acceleration [FASTA (7), BLAST (1)] with accurate profile-based dynamic programming (8), and with an automated iterative refinement of the search. As a result, the PSI-BLAST search and alignment could be even two orders of magnitude faster (9) than the corresponding Smith–Waterman (8) alignment algorithm and almost as sensitive. This is an impressive solution that clearly is one reason for the enormous popularity of PSI-BLAST. Often, everyday sequence analysis applies a two-tier approach: first a search with a reliable and fast PSI-BLAST followed by a search with programs that generate more accurate alignments but are neither fast enough nor set up for database searches such as ClustalW (10), T-Coffee (11), MAFFT (12), MUSCLE (13). Note that in the following, we use a slight deviation from the usual connotation, namely the term profile–sequence instead of sequence–profile alignment to differentiate between the query (profile) and the template/database (sequence); PSI-BLAST by this notation is a profile–sequence method.

Consensus sequences can represent families of related proteins

Protein sequences are subject to continuous evolution. Random mutations and insertions/deletions of nucleic acids within genes are source of variability of protein sequences. The pressure to maintain biological function (and/or 3D structure) constrains the range of mutations. In general, proteins can have quite dissimilar sequences and still perform the same biological function and/or have very similar 3D structure. At each sequence position, i.e. for each residue, the mutational variability can be characterized by a vector of amino acid substitution frequencies. The resulting matrix is often referred to as a *sequence profile*. The substitution frequencies are typically computed from alignments of functionally and/or structurally related proteins. In subsequent steps (iterations), such profiles are then used as the basis for aligning protein sequences (in profile-sequence and profile–profile algorithms). A consensus sequence can be thought of as a one-dimensional simplification of such a profile that, e.g. substitutes the 20-dimensional vector (for 20 amino acids) in each column (residue position) by the most frequent or most informative amino acid observed at that position. The consensus can be applied globally (to all profile columns) or locally (only to some columns) (14,15). There also exist other, more specialized techniques for generating consensus sequences (16).

Consensus sequences empower alignment methods

Consensus sequences were used early on to improve alignments (17). Initially the substitution of profiles by consensus mimicked profile–sequence alignments (14,18) (more accurately leading to consensus–sequence or sequence–consensus comparisons). Those methods tapped into fast alignment algorithms such as FASTA or BLAST. This approach is used successfully with ProDom (19) and COBBLER (14) consensus sequences. The development of fast profile–sequence alignment methods such as PSI-BLAST halted the development of sequence–consensus methods. Although BLAST-based sequence–consensus searches may be considerably faster than PSI-BLAST searches, they are thought to also be considerably less accurate. A symmetric approach of aligning a query sequence with a database of profiles (sequence–profile alignments) is used, for example, in Blocks Searcher (20) and in RPS-BLAST (6,21) to search the Blocks (22,23), PRINTS (24) and CDD (25,26) databases. Another approach is to align a query sequence with profile-derived Hidden Markov Models (HHMs) as applied by, e.g. Pfam (27) and Smart (28,29). An interesting idea suggested for PSI-BLAST searches with consensus sequences was never tested nor implemented on a larger scale (30).

Profile–profile algorithms tend to be both most sensitive and most accurate (31,32). Unfortunately, profile–profile comparisons are also much slower and more complex than heuristically accelerated sequence–sequence and profile–sequence algorithms. For this reason their application to everyday searches of large sequence databases on a typical computer workstation is not practical. Recently, an algorithm that approximates profile–profile algorithms by performing consensus–consensus alignments (16) has been published. In this article, we propose a different approximation to profile–profile comparisons in which only one profile is substituted by a consensus sequence (profile–consensus alignment). A somewhat similar approach (without heuristic speed-up) was proposed for aligning quasi-consensus sequences with HMMs (33). Consensus sequences can be derived in various ways. In one approach the raw sequences are only replaced by consensus residues 'locally', i.e. for some of the residues, e.g. the evolutionarily conserved regions (as done by the COBBLER method based on Blocks). Alternatively, one could replace the complete sequence with a consensus sequence. Here, we tested both alternatives.

Which consensus alignment is best?

Given all possible variants of using consensus sequences: which one is best? A direct comparison of existing methods may not provide the most informative answer to this question because different methods generate profiles and consensus sequences in different ways (see Supplementary Data for such a comparison). Here, we set up an experiment where we could control all the parameters to study differences between various algorithmic approaches. The same sets of multiple alignments and the same algorithms for computing consensus residues were used. Also the same alignment algorithm

(PSI-BLAST) was used to make all alignments. We compared three possible ways of using consensus sequences in alignments—aligning raw with consensus sequences (sequence-consensus), aligning only consensus sequences (consensus–consensus) and (proposed here) aligning profiles with consensus sequences (profile–consensus). In addition, we studied whether protein sequences locally enriched with consensus information performed better than simple global consensus sequences. Since the alignment of consensus sequences is as widely applicable and potentially as fast as alignment of raw sequences we have also compared it with the standard raw sequence alignment methods—PSI-BLAST and BLAST. Finally, we have provided the first comprehensive analysis for the quality of consensus sequence alignments.

We found that profile–consensus alignments outperformed other consensus sequence alignments. Notably, the profile–consensus approach most closely resembled profile–profile algorithms. The profile–consensus searches with PSI-BLAST were significantly more sensitive and specific than the original PSI-BLAST searches with raw sequences. Improvements were particularly significant for more difficult tasks such as the identification and alignment of distant structural relations between proteins. Despite the fact that the improvements were higher for more divergent relations, they were consistent even at high accuracy/low error rates for non-trivially related proteins. The improvements were very easy to achieve; no parameter used by PSI-BLAST was altered and no single line of code changed. Moreover, the consensus sequence add-on required relatively little additional CPU time. This new way of search and alignment added onto the existing PSI-BLAST program is almost as fast and easily applicable as PSI-BLAST itself.

MATERIALS AND METHODS

Generation of consensus sequences

For each test sequence used in this study, we generated the position-specific scoring matrix (PSSM) using PSI-BLAST. We used a maximum of five iterations, an e-value threshold for inclusion in PSSM of 0.001 and no query filtering [blastpgp options '–j 5 –h 0.001 –F F –Q PSSM(ASCII)']. All profiles were generated by aligning against a redundancy-reduced version of the UniProt (34) database [80% sequence identity reduction using CD-HIT (35)]. The determination of consensus amino acids was based on the ASCII PSSMs. Each original residue was replaced with the amino acid that had the highest corresponding PSSM score (highest 'target' to background frequency ratio). Three types of consensus sequences were generated: In the 'global consensus' mode, we replaced all residues by the consensus; in the 'consensus$^{top50\%}$' mode we replaced the 50% of the residues associated with most informative profile columns (highest relative entropy) by the consensus; in the 'consensus$^{low50\%}$' mode we replaced the 50% of residues associated with least informative columns with consensus residues.

Alignments

All alignments were generated using the 'blastpgp' executable in the PSI-BLAST suite of programs. All profiles (PSSMs) used for alignments were generated in the same way as profiles used for generation of consensus sequences except that a file containing the binary version of a PSSM was also stored [blastpgp options: '–j 5 –h 0.001 –F F –C PSSM(binary)']. The binary PSSM was used for a final PSI-BLAST search and alignment of the database of consensus sequences using just one iteration [blastpgp options: '–j 1 –F F –R PSSM(binary)']. For non-profile-based alignments of sequences 'blastpgp' program with default BLOSUM62 (36) scoring matrix was also used (options: '–j 1 –F F'). For comparison of performance PSI-BLAST (the same options) was used to search the corresponding database of raw sequences. For convenience of analysis the alignments of consensus sequences were translated back to 'real' sequences using a simple Perl script (Figure 1).

Evaluation of performance

There is no commonly accepted means of evaluating the performance of database search and alignment methods. One way of generating test sets of sufficient size is to compare proteins with known 3D structures because for such comparisons standards-of-truth can relatively easily be generated automatically. We assessed both the ability to identify related proteins and the ability to correctly align them based on structural alignments (below). Evolution has conserved the principle components of protein 3D structures (often misleadingly referred to as 'the fold') at higher divergence than the principle aspects of protein function. Therefore, evaluations based on structural alignments tend to put emphasis on more diverged relationships than would comparisons that are based on functional features.

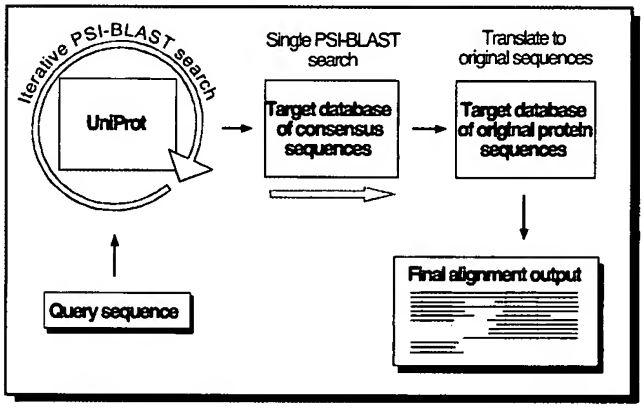


Figure 1. Sketch of consensus search. First, the PSSM for a query protein sequence is built by an iterative PSI-BLAST search over a large database of proteins sequences (such as UniProt). The resulting PSSM is then used to search and align sequences contained in a target database of consensus sequences. Finally, consensus sequence alignments are translated to alignments of the native raw protein sequences.

PSI-BLAST as the point of reference

All our evaluations used PSI-BLAST and BLAST as points of reference. The rationale was manifold. First, PSI-BLAST alone is not sufficient because there are many different ways of running PSI-BLAST, i.e. we need a point of reference in order to track our way of running PSI-BLAST. For this we explored BLAST. Second, most recent assessments of new alignment methods are compared to PSI-BLAST and/or BLAST. Since it is rather unreasonable to compare results obtained on different data sets, we cannot directly compare our results to other publications. However, the two reference points allowed for the triangulation of a comparison. Third, our major purpose was to illustrate the advantage of adding our protocol onto existing PSI-BLAST searches, i.e. PSI-BLAST is the most important point of reference for our protocol. This is because PSI-BLAST is one of the few tools that can be used for fast and accurate searching of largest sequence databases and consensus sequence alignments can be used for the same purpose.

Evaluation of search capability

We evaluated the ability to identify related proteins with SCOP (37) (release 1.69). For the assessment we omitted protein pairs from the same SCOP family (considered rather easy to recognize) and pairs that belonged to different SCOP superfamilies but to the same SCOP fold (considered too difficult for sequence alignment methods). Thus, our positives were pairs of proteins from the same SCOP superfamily while negatives were pairs of proteins from different SCOP folds.

Evaluation of alignment quality

Comparative modeling is a technique that allows the modeling of a 3D structure for a query protein Q based on a template T of experimentally known structure (38,39). In the simplest implementation comparative modeling first aligns Q and T and then copies the co-ordinates from T to model the structure of Q based on this alignment. Alignment mistakes significantly impair the quality of such models. We measured the quality of alignments implicitly, namely by assessing the quality of the comparative models originating from the alignments.

We superposed all models (represented by C_α atom coordinates) with experimentally determined 3D structures using one particular automatic method for structural superposition, namely LGA (40); this method has become one of the standards in the experiments for the Critical Assessment of Structure Prediction [CASP (41)]. First, we computed a Global Distance Test (GDT) (40) that corresponds to the largest, not necessarily continuous subset of residues superimposable within a specified distance threshold. Second, we also computed Longest Continuous Segments (LCS) (40) of residues (consecutively modeled residues) that can fit under a specified RMSD cutoff. The second measure provided us with a local alignment quality test. Note that we chose a subset of pairs (Q,T) such that for all pairs experimental structures were available; we built the model for Q using the known structure of T and assuming that Q had no known structure, but we evaluated the accuracy of the model using the experimentally known structure for Q. We reported results for two different thresholds. The first was rather stringent (2 Å); it focused on the essential core similarities between model and experiment. The second was rather relaxed (5 Å) thereby capturing more generic, coarse-grained similarities. Note that GDT computation uses the actual distance threshold while LCS uses average distance (RMSD).

Note that we assess a real-life situation in which we model structures for proteins Q that are not identical to the experimentally known structures T. This implies that the quality of a model also depends crucially on the divergence between Q and T: at high evolutionary distances, the two structures will differ so much in detail that even accurate alignments will not give as accurate models as inaccurate alignments between more closely related pairs. We accounted for this effect by structural alignments: we used the 3D alignment method MAMMOTH (42) to align the known structures of Q and T. This approximated an upper limit for what could be achieved by simplistic comparative modeling that only copied coordinates. The quality of models based on MAMMOTH alignments was also evaluated using LGA.

Data sets

We analyzed the ability to correctly identify and align related proteins on a subset of SCOP. We removed domains with discontinuous sequences, structures with missing coordinates, NMR structures, low-resolution structures (<2.5 Å) and short proteins (<50 residues). The resulting set of proteins was tailored differently for assessing search and alignment quality.

To assess search capability (homology/fold recognition), we reduced the redundancy of the sequence set so that no pair of sequences could be aligned by BLAST at e-values better than 10^{-3} (when computed on UniProt database of ~2 000 000 sequences) or at levels of sequence identity and alignment length that corresponded to HSSP-values above 0 (43,44) (whichever of the two criteria applied). This yielded a data set of 2476 sequences for which we applied an all-against-all test.

The choice of datasets for studying alignment quality was motivated by the observation that the quality of sequence alignments deteriorates rapidly below levels of around 30% pairwise sequence identity (45). In order to assess the ability of our add-on consensus approach to correctly align more distant pairs, we did not consider alignments with >30% pairwise sequence identity. Within this set of distant relatives, we monitored two different levels of alignment difficulty correlated with standard everyday uses of sequence alignment algorithms. First, we chose only those protein pairs that could be aligned by PSI-BLAST with e-values ranging from 10^{-3} to 10 when searching large public sequence databases. Second, we looked at the more difficult task of aligning protein pairs belonging to the same SCOP superfamily but different SCOP families (with e-values of up to 100 when computed on sequence unique subset of SCOP).

Those sets were composed of 1647 (set 1: most related, non-trivial pairs), and 5551 (set 2: more difficult, most diverged) protein pairs respectively. The final data sets were 'pairs non-redundant' in the following sense: no protein in any pair could be aligned with any protein from any other pair at PSI-BLAST e-values better than 1000 (calculated on UniProt database).

RESULTS

Approximation of profile–profile alignments performed best

For each alignment method tested here, we ordered all alignments of all queries by e-values. Next, we computed the cumulative number of true positive relations (same SCOP superfamily but different family; note that cases with the same superfamily and the same family were carefully filtered out from our data set to reduce redundancy) for increasing cumulative numbers of false positives (pairs of proteins with different SCOP folds). For any cumulative number of false positives (i.e. at any error rate) searching with profiles against a database of global consensus sequences yielded most true positives (profile-consensus, Figure 2A and Supplementary Data). Such a search was the closest approximation of profile–profile alignments since only one of the profiles was replaced by the corresponding consensus sequence. Replacing both profiles by consensus sequences and scoring alignments with a generic scoring matrix (BLOSUM62) did not perform as well (consensus–consensus, Figure 2A). Although this approach seemed to have some advantage over PSI-BLAST in a low error region (few false positives), the loss of some profile information for both profiles was largely detrimental. Finally, searching with a raw sequence and a generic scoring matrix against a database of consensus sequences performed worse than other consensus sequence methods but significantly better than BLAST (sequence–consensus, Figure 2A).

We also observed that global consensus sequences performed better than sequences with partial consensus information. For example, searching with profiles against consensus$^{top50\%}$ sequences (50% of the residues in most informative positions replaced by consensus) performed somewhat worse than searching against global consensus sequences (profile–consensus$^{top50\%}$, Figure 2A). Interestingly, the search with the least conserved/informative half of the residues replaced by consensus (profile–consensus$^{low50\%}$, Figure 2A) still improved performance over raw (no consensus) sequences!

Few corrupted profiles can produce many false positives with very significant scores. Alternatively, few very good profiles with many related proteins present in the database can identify them preferentially. Thus, plots of the cumulative number of true versus false positives according to alignment scores may be locally dominated by such a bias. Counting the cumulative number of true positives according to the alignment score rank obtained in each individual query search (i.e. considering the first n alignment pairs from each query) tends to reduce the bias. This test demonstrated that few outliers did not skew the results. Instead, the search against a database of global consensus sequences produced the largest number of true positives at any rank considered (Figure 2B).

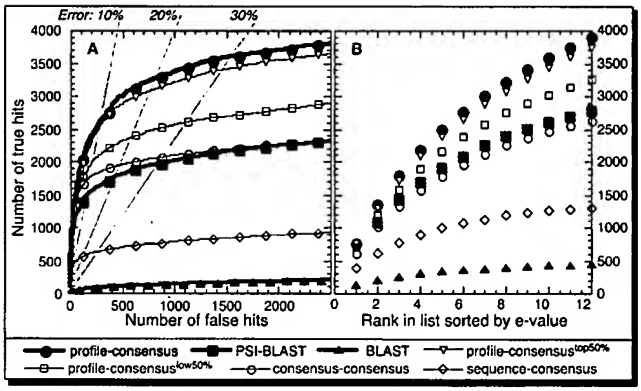


Figure 2. Consensus sequences performed better at any error rate. We compared the performance of BLAST and PSI-BLAST, with different strategies for consensus add-ons profile-consensus marked our standard approach of aligning a PSI-BLAST profile of the query against a database of consensus sequences (blue circles); profile-consensus$^{top50\%}$ aligned query profiles against a database in which only the 50% most informative residues (Methods) were replaced by consensus sequence (black inversed triangles); profile-consensus$^{low50\%}$ aligned query profiles against a database in which only the 50% least informative residues were replaced by consensus sequence (black rectangles); consensus-consensus marked BLAST-based comparisons between consensus sequences on both sides, i.e. for the database and the query (black circles); sequence-consensus were BLAST-based comparisons between native sequences on the query side and a database with consensus sequences (black diamonds). For reference, results of original sequence-based PSI-BLAST (green rectangles), and pairwise BLAST (gray triangles) are also shown. True pairs were sequences from the same SCOP superfamily (similar structure), while false ones belonged to different SCOP folds (different structure) (Methods). (**A**) Alignments (2476 sequences, all versus all) were sorted by e-values. True versus false computed over all matches found below a given e-value threshold. By construction, we excluded all pairs that were trivially related (Methods), which explained why the curves for the pairwise BLAST were so low. Profile alignments of global consensus sequences performed best. The transparent gray lines marked the levels of 10, 20 and 30% errors. For instance, at the 10% error (90% accuracy) level, the profile-based search of global consensus sequences revealed over 66% more correct relations than PSI-BLAST (global-consensus-based = 2483 true positives; PSI-BLAST = 1490). (**B**) To rule out that the improvements of consensus sequence-based searches (A) originated from few families, we counted the cumulative number of correctly classified pairs (structural similarity recognized) for the first best scoring n alignment pairs (rank n) from each query search (i.e for rank n equal 2 we looked at 4952 pairs (2 times 2476). The searches of global consensus sequences performed best at all ranks.

Little additional CPU needed for add-on. In this study, we used separate databases for iterative derivation of PSSMs (non-redundant UniProt) and for the final search and alignment ('sequence unique' SCOP; Figure 1). In this scenario, our un-optimized add-on consensus search and alignment nearly doubled the CPU time, in the following way. Five iterations of PSI-BLAST against SCOP would take about 5 s (on a single 2.8 GHz CPU with 1 GB RAM), one additional iteration of PSI-BLAST with the consensus sequence added another ~4 s. Most of the 4 s were spent on the search (3.2 s); very little additional time was needed to translate alignments of consensus sequences into 'raw' sequence alignments.

Note that we actually ran PSI-BLAST against UniProt, while we only applied the consensus addition to SCOP. The entire PSI-BLAST search against UniProt took about 5 min per query. If testing the consensus method on UniProt, we expect that this would lead to an additional 4 min of processing time.

Better alignments. We studied the alignment quality of the best performing consensus sequence search algorithm (profile–consensus) and compared it with the quality of raw sequence alignments, i.e. the original PSI-BLAST alignments. The quality was measured by assessing the 3D structure models that resulted from a simple comparative model-building strategy using these alignments. To provide a useful perspective on the results, we also evaluated 3D models obtained from structural superpositions carried out with MAMMOTH (42). We found that on average consensus sequence-based models had significantly more (not necessarily consecutive) residues in the vicinity of experimentally determined coordinates than did PSI-BLAST-based models (Table 1). This was true when measuring detailed structural similarities (stringent distance threshold of 2 Å) as well as when measuring coarse-grained structural similarities (relaxed distance threshold of 5 Å), and it was true for both levels of alignment difficulty (Table 1). However, the improvements from our add-on of consensus sequences were most significant for more difficult data sets and for more coarse-grained similarities. The comparisons with the models obtained from structural superpositions by MAMMOTH further underscored the relative significance of the gains from consensus-based searching. For example, at the threshold of 5 Å consensus-based searches increased the number of correctly modeled residues around half as much as MAMMOTH did. Surprisingly, at the stringent 2 Å threshold and the less difficult (Table 1; PSI-BLAST e-values 10^{-3}–10) consensus-aligned models had, on average, more superposed residues than the MAMMOTH models. This is rather surprising because it implies that the sequence-only alignment found a better superposition for two 3D objects than did the structure-only alignment method. The likely explanation for this puzzling finding is that MAMMOTH was optimized for the identification of structural similarities within 4 Å, i.e. a threshold more useful for more distantly related structures. In other words, structural alignments of MAMMOTH were likely not optimized for finding 'tight alignments' of closely related proteins. Nevertheless, the performance of consensus sequence-based models generated without structures was impressive in this case.

For local subsets of consecutive model residues we also found that the models resulting from the consensus alignments had longer segments of 'good quality' than did PSI-BLAST based models (Table 2). This was true for a more stringent RMSD threshold of 2 Å as well as for a more relaxed threshold of 5 Å. Again, the improvements from our add-on of consensus sequences were most

Table 1. Consensus sequences improve the global quality of structural models*

	<2 Å (C_α distance)			<5 Å (C_α distance)		
	PSI-BLAST	PROFILE-CONSENSUS	MAMMOTH	PSI-BLAST	PROFILE-CONSENSUS	MAMMOTH
SCOP superfamily, only	15.8 (±0.2)	**18.1** (±0.2)	19.6 (±0.2)	22.6 (±0.2)	**27.2** (±0.3)	35.2 (±0.3)
PSI-BLAST e-values 10^{-3}–10	34.7 (±0.4)	**38.3** (±0.5)	36.5 (±0.4)	49.1 (±0.5)	**55.2** (±0.6)	58.0 (±0.5)

*For each protein in our data sets (query Q), we aligned a similar protein in the PDB (template T) and used the experimental structure of T to model the structure for Q by simply copying the C_α backbone of T onto Q according to the alignment provided. Since for all Qs in our experiment the correct answer was known (all Qs had known structure), we could then assess how accurate the model was by superposing the model and the known structure. For this superposition, we used the structural alignment method LGA. Here, the measure of accuracy was the percentage of C_αs that were closer to the real structure than some distant cutoff (<5 Å for the three rightmost columns, and <2 Å for columns 2–4). Note that the set of residues below a distance threshold was not necessarily consecutive in sequence. We compared the consensus sequence-based approach with that of the regular PSI-BLAST. The data for MAMMOTH was generated by optimally superposing the structures of Q and T without considering their sequences. In principle, this approximated an upper threshold for performance (Results). The two rows distinguished different data sets corresponding to different levels of alignment difficulty: '***SCOP superfamily only***' were pairs of proteins that fell into different SCOP families and into the same SCOP superfamily (coarse-grained structural relation), while '***PSI-BLAST e-values 10^{-3}–10***' were pairs of proteins with similar structure that fell into the corresponding interval of sequence similarity. Note that both rows reflected the performance for 'non-trivial' tasks. Standard errors are given in parentheses.

Table 2. Consensus sequences improve the local quality of structural models*

	<2 Å (C_α RMSD)			<5 Å (C_α RMSD)		
	PSI-BLAST	PROFILE-CONSENSUS	MAMMOTH	PSI-BLAST	PROFILE-CONSENSUS	MAMMOTH
SCOP superfamily, only	14.4 (±0.1)	**15.7** (±0.2)	16.5 (±0.2)	23.8 (±0.3)	**27.9** (±0.3)	35.3 (±0.3)
PSI-BLAST e-values 10^{-3}–10	26.8 (±0.4)	**28.0** (±0.4)	26.1 (±0.4)	51.6 (±0.6)	**58.4** (±0.7)	62.8 (±0.7)

*Data sets identical to those as in Table 1; the difference is that accuracy is now measured by considering a single sequence-consecutive segment in the model that falls below a certain distance threshold. The longest consecutive segments were identified by the program LGA. Note that thresholds reflect cutoffs in terms of C_α RMSD, i.e. the distance averaged over the entire segment. In contrast, the values in Table 1 reflect actual C_α thresholds for spatial distances.

significant for more difficult data sets and for more coarse-grained similarities.

DISCUSSION

Here we demonstrated that both the search and alignment quality of PSI-BLAST can easily be improved without having to alter the code. Performance improved substantially with simply replacing the last iteration of the standard PSI-BLAST search against a database of raw sequences with a search against a database of consensus sequences. The improvements were most significant for non-trivial tasks such as the identification (Figure 2) and alignment of distant structural similarities. All improvements translated directly into better initial models for comparative modeling (Tables 1 and 2).

The analysis provided a worst-case scenario for the performance of consensus sequences resulting from simply piggybacking a new idea (usage of consensus sequences directly for the alignment) onto an old method (PSI-BLAST). We neither altered gap penalties (11 for opening and 1 for extension), nor substitution matrices, nor any other parameter optimized for raw rather than consensus sequences. Preliminary tests (data not shown) indicated that consensus sequence-based searches did not change the robustness/sensitivity with respect to such parameters. We also found that using the most frequent amino acid type at each position instead of the amino acid with maximal PSSM score did not reduce the gain significantly. On the other hand, the adverse consequence of not optimizing any of the PSI-BLAST parameters was that searching a database of consensus sequences took almost four times as long as searching a comparable database of raw sequences (~3.2 versus ~0.8 s on a non-redundant SCOP). Lately, we have realized that it was largely due to using parameters such as thresholds for extending hits (high-scoring residue words), triggering gapped alignments and gap penalty values themselves that were not optimal for consensus sequences (our preliminary results indicate that raising the threshold for extending hits by about 20% almost doubles the speed and affects the sensitivity negligibly). Those details, as well as the scoring matrix, remain to be optimized for the particular concept of consensus sequences.

To generate global consensus sequences, we replaced each amino acid in the template by the amino acid that scored highest in the associated column of a profile PSSM produced by a standard PSI-BLAST search. Thereby, we maximized the self-score of the resulting consensus sequence with respect to its PSSM. As a consequence, any two proteins having similar profiles are also likely to have a higher alignment score when consensus sequences are aligned. Our results suggest that the corresponding change of the alignment score for unrelated proteins was considerably smaller. Surprisingly, replacing only the least informative half of all residues by consensus also improved performance (profile-consensus$^{low50\%}$, Figure 2). This may suggest that even weakly or non-conserved positions are associated with specific constraints on random amino mutations that can be utilized to detect similarities.

The best performance of profile–consensus search was achieved when the profile that was used to generate the consensus sequence was obtained in the same way as the profile used for the alignment scoring. For example, when the profile used to compute the consensus was obtained after fewer PSI-BLAST iterations, performance deteriorated. Improving the searches through consensus databases that apply more involved ways of using consensus sequences such as ProDom and COBBLER may therefore require one to search with the same type of scoring profiles that was used to generate the database in the first place. Unfortunately, the algorithms used for their creation are considerably more involved and more time consuming. In contrast, our add-on protocol is very simple. The global consensus sequences can be generated easily from PSI-BLAST ASCII matrices. The optimal search of such database requires similarly easily obtainable PSI-BLAST binary profiles. Any PSI-BLAST user could easily accomplish this. However, the generation of a large consensus database is computationally costly. Therefore, we decided to provide an up-to-date consensus sequence version of Swiss-Prot (46) and PDB (47) through our website (http://www.rostlab.org/services/consensus/). We plan to provide consensus sequences for the entire UniProt in the near future. We have also provided a simple Perl program for translating PSI-BLAST ASCII matrices into consensus sequences. In addition, for the convenience of users we have provided a script for the conversion of aligned consensus sequences into the corresponding alignments of real sequences. We have also made profile–consensus searches available through the PredictProtein server (48) (http://predictprotein.org).

Our results suggested that sequence–profile method (i.e. methods that search database of profiles with a sequence) such as IMPALA and the methods used to search CDD (25,26) might also benefit from mimicking profile–profile alignments through searching database of profiles with a consensus sequence (consensus–profile alignment). Similarly, methods that use sequences to search HMM-derived profile databases such as in Pfam and SMART might also improve performance by replacing a raw query sequence with a consensus sequence as proposed in this manuscript, although the HMM-derived consensus sequences may be more appropriate (33). Finally, it is also likely that methods using bidirectional profile–sequence/sequence–profile scoring (49,50) will benefit from using profile–consensus/consensus–profile approach.

One advantage other than improved performance is that consensus sequence-based alignments are likely less sensitive to sequencing errors. This may be particularly appealing in the age of massive sequencing efforts that grind up indiscriminately what is found in oceans, soils and polluted environments. Finally, it remains to be shown that the advantage of using consensus sequence-based searches for the identification and alignment of remote structural similarities between proteins will hold more generally, e.g. for the nucleotide sequences, and

for the usage of with other alignment algorithms, such as ClustalW or T-Coffee.

One consequence of our improvements was that the consensus sequence-based alignment profiles were both more diverse and more accurate than those generated by the ordinary PSI-BLAST. Prediction methods that use alignment profiles, such as those predicting aspects of protein structure, tend to improve proportionally with better profiles (51–54). It is therefore reasonable to assume that our consensus sequence add-on to PSI-BLAST will clearly boost the performance of downstream methods for the prediction of protein structure and function.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Thanks to Henry Bigelow, Eileen Guilfoyle, and Kazimierz Wrzeszczynski for reading and commenting on the manuscript and to Guy Yachdav for help in setting up the server. Thanks to all those who develop alignment tools and make them available to the public. In particular thanks to people developing and providing support for the PSI-BLAST program. Thanks to Adam Zemla for providing LGA and to the creators and developers of MAMMOTH. The support for the work of DP and BR and funding to pay the Open Access publication charge was provided by the grant U54-GM074958-01 to the Northeast Structural Genomics consortium (NESG) from the Protein Structure Initiative (PSI) of the National Institutes of Health (NIH) and by the grant RO1-LM07329-01 from the National Library of Medicine (NLM). Last but not least, thanks to all of those who maintain excellent databases and to all experimentalists who enabled this analysis by making their data publicly available.

Conflict of interest statement. None declared.

REFERENCES

1. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
2. Pietrokovski,S. (1996) Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Res.*, **24**, 3836–3845.
3. Rychlewski,L., Jaroszewski,L., Li,W. and Godzik,A. (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci.*, **9**, 232–241.
4. Sadreyev,R. and Grishin,N. (2003) COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J. Mol. Biol.*, **326**, 317–336.
5. Yona,G. and Levitt,M. (2002) Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. *J. Mol. Biol.*, **315**, 1257–1275.
6. Altschul,S., Madden,T., Schäffer,A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D. (1997) Gapped Blast and PSI-Blast: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
7. Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA*, **85**, 2444–2448.
8. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
9. Altschul,S.F. and Koonin,E.V. (1998) Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases. *Trends Biochem. Sci.*, **23**, 444–447.
10. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
11. Notredame,C., Higgins,D.G. and Heringa,J. (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.
12. Katoh,K., Misawa,K., Kuma,K. and Miyata,T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, **30**, 3059–3066.
13. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
14. Henikoff,S. and Henikoff,J.G. (1997) Embedding strategies for effective use of information from multiple sequence alignments. *Protein Sci.*, **6**, 698–705.
15. Sigrist,C.J., Cerutti,L., Hulo,N., Gattiker,A., Falquet,L., Pagni,M., Bairoch,A. and Bucher,P. (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief. Bioinformatics*, **3**, 265–274.
16. Merkeev,I.V. and Mironov,A.A. (2006) PHOG-BLAST—a new generation tool for fast similarity search of protein families. *BMC Evol. Biol.*, **6**, 51.
17. Patthy,L. (1987) Detecting homology of distantly related proteins with consensus sequences. *J. Mol. Biol.*, **198**, 567–577.
18. Sonnhammer,E.L. and Kahn,D. (1994) Modular arrangement of proteins as inferred from analysis of homology. *Protein Sci.*, **3**, 482–492.
19. Servant,F., Bru,C., Carrere,S., Courcelle,E., Gouzy,J., Peyruc,D. and Kahn,D. (2002) ProDom: automated clustering of homologous domains. *Brief. Bioinformatics*, **3**, 246–251.
20. Henikoff,S. and Henikoff,J.G. (1994) Protein family classification based on searching a database of blocks. *Genomics*, **19**, 97–107.
21. Marchler-Bauer,A., Anderson,J.B., Cherukuri,P.F., DeWeese-Scott,C., Geer,L.Y., Gwadz,M., He,S., Hurwitz,D.I., Jackson,J.D. *et al.* (2005) CDD: a Conserved Domain Database for protein classification. *Nucleic Acids Res.*, **33**, D192–196.
22. Henikoff,J.G., Greene,E.A., Pietrokovski,S. and Henikoff,S. (2000) Increased coverage of protein families with the blocks database servers. *Nucleic Acids Res.*, **28**, 228–230.
23. Henikoff,S., Henikoff,J.G. and Pietrokovski,S. (1999) Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics*, **15**, 471–479.
24. Attwood,T.K., Bradley,P., Flower,D.R., Gaulton,A., Maudling,N., Mitchell,A.L., Moulton,G., Nordle,A., Paine,K. *et al.* (2003) PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res.*, **31**, 400–402.
25. Schäffer,A.A., Wolf,Y.I., Ponting,C.P., Koonin,E.V., Aravind,L. and Altschul,S.F. (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics*, **15**, 1000–1011.
26. Marchler-Bauer,A., Panchenko,A.R., Shoemaker,B.A., Thiessen,P.A., Geer,L.Y. and Bryant,S.H. (2002) CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res.*, **30**, 281–283.
27. Finn,R.D., Mistry,J., Schuster-Bockler,B., Griffiths-Jones,S., Hollich,V., Lassmann,T., Moxon,S., Marshall,M., Khanna,A. *et al.* (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, **34**, D247–251.
28. Letunic,I., Copley,R.R., Pils,B., Pinkert,S., Schultz,J. and Bork,P. (2006) SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res.*, **34**, D257–260.
29. Schultz,J., Milpetz,F., Bork,P. and Ponting,C.P. (1998) SMART, a simple modular architecture research tool:

identification of signaling domains. *Proc. Natl. Acad. Sci. USA*, **95**, 5857–5864.
30. Thelen,M.P., Venclovas,C. and Fidelis,K. (1999) A sliding clamp model for the Rad1 family of cell cycle checkpoint proteins. *Cell*, **96**, 769–770.
31. Kryshtafovych,A., Venclovas,C., Fidelis,K. and Moult,J. (2005) Progress over the first decade of CASP experiments. *Proteins*, **61** Suppl 7, 225–236.
32. Rychlewski,L. and Fischer,D. (2005) LiveBench-8: the large-scale, continuous assessment of automated protein structure prediction. *Protein Sci.*, **14**, 240–245.
33. Kahsay,R.Y., Wang,G., Gao,G., Liao,L. and Dunbrack,R. (2005) Quasi-consensus-based comparison of profile hidden Markov models for protein sequences. *Bioinformatics*, **21**, 2287–2293.
34. Apweiler,R., Bairoch,A., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R. *et al.* (2004) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **32**, D115–119.
35. Li,W., Jaroszewski,L. and Godzik,A. (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, **17**, 282–283.
36. Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA*, **89**, 10915–10919.
37. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
38. Fiser,A., Feig,M., Brooks,C.L. 3rd and Sali,A. (2002) Evolution and physics in comparative protein structure modeling. *Acc. Chem. Res.*, **35**, 413–421.
39. Ginalski,K. (2006) Comparative modeling for protein structure prediction. *Curr. Opin. Struct. Biol.*, **16**, 172–177.
40. Zemla,A. (2003) LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res.*, **31**, 3370–3374.
41. Moult,J., Fidelis,K., Rost,B., Hubbard,T. and Tramontano,A. (2005) Critical assessment of methods of protein structure prediction (CASP)—round 6. *Proteins*, **61** Suppl 7, 3–7.
42. Ortiz,A.R., Strauss,C.E. and Olmea,O. (2002) MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci.*, **11**, 2606–2621.
43. Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Eng.*, **12**, 85–94.
44. Sander,C. and Schneider,R. (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.
45. Baker,D. and Sali,A. (2001) Protein structure prediction and structural genomics. *Science*, **294**, 93–96.
46. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
47. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
48. Rost,B., Yachdav,G. and Liu,J. (2004) The PredictProtein server. *Nucleic Acids Res.*, **32**, W321–326.
49. Kelley,L.A., MacCallum,R.M. and Sternberg,M.J. (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.*, **299**, 499–520.
50. Przybylski,D. and Rost,B. (2004) Improving fold recognition without folds. *J. Mol. Biol.*, **341**, 255–269.
51. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.
52. Przybylski,D. and Rost,B. (2002) Alignments grow, secondary structure prediction improves. *Proteins*, **46**, 197–205.
53. Rost,B. (2001) Review: protein secondary structure prediction continues to rise. *J. Struct. Biol.*, **134**, 204–218.
54. Rost,B. and Sander,C. (1993) Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.*, **232**, 584–599.

Published online before print January 28, 2003, 10.1073/pnas.252753899
PNAS | **February 18, 2003** | vol. 100 | no. 4 | **1679-1684**

◄ Previous Article | Table of Contents | Next Article ►

Biophysics

Associative memory Hamiltonians for structure prediction without homology: α/β proteins

Corey Hardin[†], Michael P. Eastwood[‡], Michael C. Prentiss[‡], Zadia Luthey-Schulten[†,§], and Peter G. Wolynes[‡,¶]

[†] Center for Biophysics and Computational Biology and [§] School of Chemical Sciences, University of Illinois, 600 South Mathews Avenue, Urbana, IL 61801; and [‡] Department of Chemistry and Biochemistry, University of California at San Diego, 9500 Gilman Drive, La Jolla, CA 92093

Contributed by Peter G. Wolynes, December 10, 2002

This Article
- Abstract FREE
- Full Text (PDF)
- Alert me when this article is cited
- Alert me if a correction is posted
- Citation Map

Services
- Similar articles in this journal
- Similar articles in ISI Web of Science
- Similar articles in PubMed
- Alert me to new issues of the journal
- Add to My File Cabinet
- Download to citation manager
- Cited by other online articles
- Search for citing articles in: ISI Web of Science (9)
- Request Copyright Permission

Google Scholar
- Articles by Hardin, C.
- Articles by Wolynes, P. G.
- Articles citing this Article
- Search for Related Content

PubMed
- PubMed Citation
- Articles by Hardin, C.
- Articles by Wolynes, P. G.

Abstract

We describe a method for predicting the structure of α/β class proteins in the absence of information from homologous structures. The method is based on an associative memory model for short to intermediate range in sequence contacts and a contact potential for long range in sequence contacts. The coefficients in the energy function are chosen to maximize the ratio of the folding temperature to the glass transition temperature. We use the resulting optimized model to predict the structure of three α/β protein domains

▲ Top
▪ Abstract
▼ Introduction
▼ Materials and Methods
▼ Constrained Self-Consistent...
▼ Results and Discussion
▼ Conclusion
▼ Appendix

▼ References

ranging in length from 81 to 115 residues. The resulting predictions align with low rms deviations to large portions of the native state. We have also calculated the free energy as a function of similarity to the native state for one of these three domains, and we show that, as expected from the optimization criteria, the free energy surface resembles a rough funnel to the native state. Finally, we briefly demonstrate the effect of roughness in the energy landscape on the dynamics.

Introduction

▲ Top
▲ Abstract
• Introduction
▼ Materials and Methods
▼ Constrained Self-Consistent...
▼ Results and Discussion
▼ Conclusion
▼ Appendix
▼ References

The rapid expansion of the protein sequence databases brought about by, among other things, the various genome sequencing projects has intensified interest in the problem of protein structure prediction. In recent years, there has been much progress toward the goal of predicting protein structure from sequence. Indeed, prediction is now almost routine for sequences with a moderate degree of homology (typically 30-50% sequence identity) to a protein of known structure (1). When homologous structures are not available prediction is more difficult, but even here, there has been much progress (2). Following Anfinsen's (3) thermodynamic hypothesis, algorithms for *ab initio* prediction typically involve the minimization of some model energy function. Although several energy functions (4-7) have been successful in generating low-resolution structures most suffer from an incomplete correlation between the energy and the quality of the prediction (2, 8, 9).

Advancing in parallel with techniques for structure prediction has been the theoretical understanding of the protein folding reaction itself. The large number of degrees of freedom needed to characterize a folding protein chain naturally leads to the adoption of a statistical characterization of the protein energy landscape (10). Such a characterization reveals that the ability of a protein to reliably find its native state among the exponentially large number of conformations is caused by the topography of the landscape. Inter-residue contacts that appear in the native state are, on average, more stabilizing than random contacts so that both the energy and entropy drop as the protein approaches the native state, and the landscape resembles a rough funnel. Bryngelson and Wolynes (11) have termed this property of the landscape the "principle of minimal frustration." Model energy functions for structure prediction must also be minimally frustrated, and for the same reason, to overcome the multiple minima problem. This insight, that the essential physics of folding is contained in the requirement of minimal frustration, and not so much in the detailed form of the interaction potentials, is at the heart of a fruitful interaction between analytical models of the folding reaction and the development of practical methods of structure prediction.

We have developed a series of models (7, 12-14) based on associative memory energy functions. By formulating a quantitative version of the principle of minimal frustration, we have optimized the coefficients in our models to achieve a minimally frustrated landscape and have shown that the resulting energy function can successfully predict low-resolution structures in the absence of homology information for α-helical proteins. Moreover, we are able to calculate the free energy as a function of similarity to the native state and thereby quantify the success of the optimization procedure in achieving a funneled landscape. Here we further develop this approach and report the successful *ab initio* prediction of α/β proteins.

The organization of this article is as follows. First, we describe a number of changes we have made to the energy function and the optimization procedure needed to adapt it to α/β structures. We then describe the results of prediction runs on members of the set of the proteins used to optimize the model and on three α/β proteins not related to any of the training proteins. Finally, we discuss the full free energy surface of one of the test proteins as a function of similarity to the native state, and we briefly discuss the dynamics of the model.

Materials and Methods

▲ Top
▲ Abstract

Potential Function for α/β Structures.

The structure prediction protocol reported here is a modification of the one described in detail in refs. 7, 9, and 14 and is based on the associative memory energy functions first introduced by Friedrichs and Wolynes (12). For completeness, we will briefly review the main features of the earlier work. The energy of a protein conformation is a function of the similarity of the set of pair distances associated with that conformation to the aligned pair distances in a database of known protein structures. For *de novo* protein prediction, the database contains only proteins that are globally unrelated to the target sequence.

We use a reduced representation of the chain consisting of C_α, C_β, and O atoms. For short to intermediate sequence separations the conformational energy is given by an associative memory energy function: $(V_{AM} = -\sum_\mu^n \sum_{i<j}^N \gamma(P_i,P_j,P_{i'}^\mu,P_{j'}^\mu,(j-i),SS_{i'},SS_{j'})\Theta(r_{ij} - r_{i'j'}^\mu)$. The coefficients, γ, weight the different types of interactions and are functions of the chemical properties (P) of the amino acids i and j, their sequence separation, the identity of residues i' and j', in the database protein μ (13) and the secondary structure, $SS_{i'}$ and $SS_{j'}$, of the database residues. We use a previously described sequence-structure **alignment** algorithm to associate the ij and $i'j'$ pairs (15). We use a four-letter code for the amino acid properties, hydrophobic, polar, acid/hydrophilic, and base, along with three sequence proximity classes, short ($j - i \leq 4$), intermediate ($5 \leq j - i \leq 12$), and tertiary ($j - i \geq 13$). In contrast to previous work we do not allow the interaction between residues to depend on their order in the chain and set $\gamma_{i,j,i',j'} = \gamma_{j,i,j',i'}$.

At large sequence separations the conformational energy is given by a simple contact potential with the form:

$$V_{long}(P_i, P_j, r_{ij}) = \sum_{k=1}^{5} C_k(N)\gamma(P_i, P_j, k)U_k(r_{ij}), \qquad [1]$$

where the $U(r_{ij})$ are designed to approximate square-well potentials about the distance ranges 4.5-6.5 Å, 6.5-8.5 Å, 8.5-10.5 Å, 10.5-12.5 Å, and 12.5-15.0 Å. To increase the discriminatory power of the tertiary potential, we have increased the number of wells since our earlier work (14). $C_k(N)$ is a scaling term that accounts for the variation in the number of contacts in each of the five wells in native protein structures of N residues in length. It has the form $a_kN/(1.0 + b_kN)$. The values of the parameters are given in Table 1.

Table 1. Parameters of the contact potential scaling term, $C_k(N)$

View this table:
[in this window]
[in a new window]

The formation of β-stranded structures critically depends on the stabilization from interstrand hydrogen bonding, a feature absent from helical proteins. For this reason, we have added several new patterns of interactions to our previous hydrogen bond term:

$$V(ij)_{HB} = -\lambda_{HB}(|\, i-j\,|)\exp\left[\frac{-(r_{ij}^{ON} - \langle r^{ON}\rangle)^2}{2\sigma_{NO}^2} - \frac{(r_{ij}^{OH} - \langle r^{OH}\rangle)^2}{2\sigma_{HO}^2}\right], \qquad [2]$$

where r_{ij}^{ON} denotes the distance from the carbonyl oxygen on residue i to the nitrogen on residue j, and r_{ij}^{OH} denotes the distance from the oxygen on residue i to the H-bonded hydrogen on residue j. First, in an effort to foster the cooperative formation of regular secondary structure elements, we added an additional dependence on the presence or absence of hydrogen bonds between nearby residues:

$$V(ij)_{HB} = -\lambda_1(|\,j-i\,|)\theta_{ij} - \lambda_2(|\,j-i\,|)\theta_{ij}\theta_{ji} - \lambda_3(|\,j-i\,|)\theta_{ij}\theta_{j,i+2},$$

where the θ functions are exponentials of the form given in Eq. 2. The λ_2 term gives an additional stabilization to an antiparallel β hydrogen bonding, and the λ_3 does the same for parallel β patterns. The dependence on $|j-i|$ indicates that the coefficients are set separately for each proximity class. The final values of the coefficients were optimized to maximize the free energy difference between the native and unfolded states as described (16) and are listed in Table 2.

Table 2. Coefficients of the hydrogen bond term

View this table:
[in this window]
[in a new window]

The registry of β strands is often poorly encoded by the Hamiltonian using only a four-letter code. To correct this we have made use of a suggestion by Regan and others (17, 18) that β secondary structures are stabilized by specific pair interactions as well as amino acid preferences. To account for these interactions, we have introduced a sequence dependence to the nonadditive coefficients λ_2 and λ_3:

$$\lambda_2(a_i, a_j) = \lambda_2 - \alpha_1 \ln P_{anti}(a_i) + \alpha_1 \ln P_{anti}(a_j)$$

$$+0.5\alpha_2(|\,j-i\,|)\ln P_{HB}(a_i, a_j) - 0.25\alpha_3(|\,j-i\,|)\{\ln(P_{NHB}(a_{i+1}, a_{j-i}) + \ln P_{NHB}(a_{i-1}, a_{j+1}))\}$$

$$\lambda_3(a_i, a_j) = \lambda_3 - [\alpha_4 \ln P_{par}(a_{i+1}) + \alpha_4 \ln P_{par}(a_j)$$

$$+\alpha_5(|\,j-i\,|)\ln P_{par}(a_{i+1}, a_j)].$$

The probabilities, P, for amino acids to be in particular secondary structures were computed by using a database of well-resolved x-ray structures as follows:

$$P_{anti}(a_i) = (N^{a_i}_{anti}/N_{anti})/(N^{a_i}/N)$$

$$P_{par}(a_i) = (N^{a_i}_{par}/N_{par})/(N^{a_i}/N)$$

$$P_X(a_i, a_j) = (N^X_{a_i,a_j}/N^X_{pairs})/[\,N^X_{a_i}N^X_{a_j}/(N^X_{pairs})^2\,],$$

where X can refer to hydrogen-bonded, nonhydrogen-bonded, or parallel pairs as defined by Regan and coworkers (17). The final values of the probabilities are in good agreement with the experimental values reported by Regan and coworkers (17) and the calculations of Wouters and Curmi (18). The coefficients, α_i, were optimized as above and are given in Table 2. The total hydrogen bond potential, V_{HB}, is the sum over the contribution from each pair, $V(ij)_{HB}$.

The hydrogen bond term as defined is fairly narrow; i.e., even relatively small deviations from ideal β-sheet geometry lead to a large loss of hydrogen bond energy. This is desirable from the point of view of reproducing the geometry of secondary structure elements accurately; however, it is disadvantageous in the search for a globally correct fold to have only such a

strict definition of a hydrogen bond, because at temperatures where many hydrogen bonds form the barriers to breaking them will be large, leading to slow dynamics. In the spirit of making a funneled (rather than golf course-shaped) landscape, we introduce a further term to the energy function intended to encourage β strands to line up in a roughly parallel or antiparallel manner even at temperatures where the hydrogen bonding has not fully set in. This potential is based on C_α positions and gives a reduction in the total energy if when residues i and j are in contact $i + 4$ and $j + 4$ (parallel, P) or $i + 4$ and $j - 4$ (antiparallel, AP) are also in contact. The P and AP contacts are allowed different weights, and the AP term is itself split into two distance classes (AP and APH) to allow different weights for putative β-hairpins. This term is thus a sum of three parts,

$$V_{P-AP} = -\gamma_{APH} \sum_{i=1}^{N-13} \sum_{j=i+13}^{\min(i+20,N)} v_{ij} v_{i+4,j-4} - \gamma_{AP} \sum_{i=1}^{N-21} \sum_{j=i+21}^{N} v_{ij} v_{i+4,j-4} - \gamma_P \sum_{i=1}^{N-17} \sum_{j=i+13}^{N-4} v_{ij} v_{i+4,j+4},$$

where $v_{ij} = 1/2(1 + \tanh[7(8 - r_{ij})])$. The coefficients, γ, are all set to 0.4ϵ.

Finally, we have introduced two new features to the energy function that enable us to take advantage of additional information that may be available about a target sequence before predicting its structure. To the Ramachandran potential described in ref. 14, V_{rama}, we have added two wells centered at dihedral angels appropriate for α-helices and β-sheets, respectively. The coefficients on these wells can then be used to provide the option of biasing the protein backbone to its predicted secondary structure:

$$V_i^{bias}(\varphi_i, \psi_i) = \lambda_i^\alpha \exp(-419.0\{[\cos(\phi_i + 0.995) - 1]^2 + [\cos(\psi_i + 0.820) - 1]^2\}) + \lambda_i^\beta \exp(-15.398\{[\cos(\phi_i + 2.25) - 1]^2 + [\cos(\psi_i - 2.16) - 1]^2\}).$$

For the set of test proteins discussed here the target sequence was submitted to the JPRED (19) secondary structure prediction server, and λ_i^α was set to 2.0ϵ for residues predicted to be helical and zero for all other residues. Similarly, λ_i^β was set to 2.0ϵ for those residues predicted to be β and zero otherwise. It has also been shown (9, 20) that averaging interactions over homologous sequences can improve the free energy surface of structure prediction energy functions. In several of the runs discussed below, we have done a multiple sequence **alignment** of top scoring hits from a PSI-BLAST (21) search with the target sequence and computed separate potentials for each sequence (including the target) in the **alignment**. Molecular dynamics on the target sequence is then performed with the average force.

The total energy function also includes terms for amino acid chirality, an excluded volume term, and a combination of harmonic terms and SHAKE (22) constraints that maintain the planarity of the peptide bond, and appropriate bond lengths, and bond angles. The coefficients for these terms are the same as used previously. The full, modified associative memory and contact (AMC) energy function, including the backbone, is:

$$V_T = -(V_{AM} + V_{long} + \lambda_{\phi\psi} V_{\phi\psi} + \lambda_{HB} V_{HB} + \lambda_\chi V_\chi + \lambda_{EV} V_{EV} + \lambda_{Harm} V_{Harm} + V_{P-AP}).$$

We define a reduced temperature as $T^* = k_B T/\epsilon$. Here ϵ is one-quarter of the native state energy per residue averaged over the training in the following way:

$$\epsilon = \left(\frac{E_{AM+C}^{Native}}{4N} \right).$$

With this choice of units, the folding temperature is typically near $T^* = 1.0$.

Constrained Self-Consistent Optimization

The parameters in the AMC energy function should be chosen to give good discrimination between the native state and typical unfolded states at intermediate temperatures and to minimize the presence of local minima that can slow the search through conformational space. The minimal requirement for rapid folding of a target sequence is a sufficiently large ratio of the stability gap, δE_S, the gap in energy between the native state and the average energy of the ensemble of non-native states and the variance in energy of the unfolded states ($\delta E_S/\Delta E$). The stability gap is related to the folding temperature, T_F, and the variance is related to typical depth of a local minimum, and thus to the glass transition temperature, T_G (10). Maximizing the ratio of the stability gap to the variance can be shown to be equivalent to maximizing the ratio of the folding temperature to the glass transition temperature (23).

As described, we enforce a set of constraints on the contribution to the mean energy of the globules from each proximity class, and we enforce roughly equal transition temperatures in each proximity class by constraining the variance in each class. The details of the optimization functional are contained in Hardin *et al.* (14).

To determine the optimal set of parameters, we choose a training set of 14 α/β proteins and generate a set of unfolded conformations via a constant temperature molecular dynamics simulation. The full set of 14 training proteins and their associated memories are discussed in the *Appendix*. To generate the initial set of decoys, we used an energy function that was optimized for an α-helical training set (7). Once the optimum set of parameters is chosen for a particular ensemble of unfolded states that energy function is used to generate a new set of decoys, and the procedure is iterated until self-consistency (13). The collapse temperature is related to the mean energy of the unfolded states and can vary among the members of the training set. To ensure that the globules for each training set protein come from roughly equivalent portions of phase diagram, we constrain the unfolded states to have a given degree of similarity to the native state. This is measured by the fraction of native contacts, or Q:

$$Q = \frac{2}{(N-1)(N-2)} \sum_{i<j-2} \exp\left[-\frac{(r_{ij} - r_{ij}^N)^2}{2\sigma_{ij}^2} \right].$$

The unfolded ensembles were constrained to have a Q of 0.3. The constraint procedure is described in ref. 9.

Results and Discussion

Once the optimized energy function is obtained we minimize it by using simulated annealing via molecular dynamics. Table 3 shows the results of simulations on each of the 14 training set proteins, using just the optimized energy function, i.e., without applying the bias to predicted secondary structure or the average over a multiple sequence **alignment**. The database of known structures from which the AMC potential is calculated was constructed by deliberately excluding any proteins with structural similarity to the corresponding training set protein. Thus scaffold proteins have global rms deviations (rmsd) that are generally >9 Å (9). There are two points about the results

on the training set proteins. First, the best structure obtained in the simulation is frequently more native-like, as measured by Q, than anything in the database. This demonstrates the ability of the AMC potential to reconstruct the members of the training set by generalizing from the partial structural similarities contained in the **alignments** to globally unrelated structures. Finally, there is a general decline in quality of the predictions as the length of the target sequence increases. The potential is most effective for sequences <90 residues long. For most of the training proteins beyond that length, the best structure obtained is somewhat inferior to the best input structure from the point of view of the Q measure. This may indicate a generic size dependence of the potentials that is not accounted for in our model. The use of the bias to predicted secondary structure and the averaging over sequence homologs should generally improve the performance of the potential. To test this expectation, we have conducted five simulations each on proteins 2acy and 3chy with the augmented energy function. In the case of 2acy, the Q_{best} structure is improved compared with the previous results; however, for 3chy it is unchanged. We have used the augmented potential for all of the test set simulations, discussed below.

Table 3. Results of simulated annealing on training set proteins

View this table:
[in this window]
[in a new window]

To test the optimized potential, we choose three protein domains from the critical assessment structure prediction 4 experiment. The test set proteins are domain 1C from FtsA (Protein Data Bank code 1E4F, residues 86-166), residues 200-309 of *Streptococcus mutans* Pyrophosphatase, and the N-terminal 115 residues of the human XRCC4 DNA repair protein. The highest Q to any member of the associative memory database used for each of these targets, Q_{mem}, is given in Table 4. We have also included the structural **alignment** from the combinatorial extension (CE) program of Shindyalov and Bourne (24). CE finds the **alignment** of two proteins that maximizes the structural overlap. Table 4 reports the length of the **alignment**, the number of residues contained in gaps in that **alignment**, the rmsd of the **alignment**, and a statistical score, Z, which is a function of the difference between the **alignment** score and the distribution of scores associated with random **alignments**. $Z > 4.0$ typically denotes a strong structural similarity; $3.7 \leq Z \leq 4.0$ represents a more ambiguous structural assignment (24). The low Z values, taken together with the low Qs, demonstrate that the three test set proteins are structurally unrelated to the database proteins. Table 5 indicates that the three test set proteins are also unrelated to any of the training set proteins, and so constitute a test of the AMC potential's performance on an unknown target.

Table 4. Structural relationship of test set proteins to database proteins

View this table:
[in this window]
[in a new window]

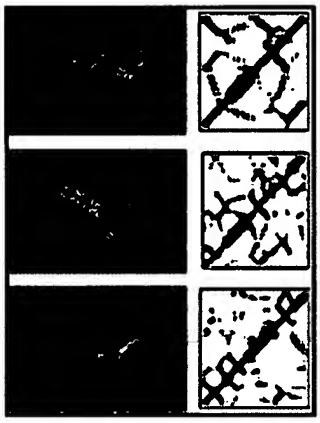
Table 5. Structural relationship between test set and training set proteins

View this table:
[in this window]
[in a new window]

The simplest gauge of the success of a prediction is the global superposition of the predicted and correct structure. Fig. 1 illustrates such a superposition for the best Q structure (Q_{best}) encountered during the simulations on each of the test set proteins. Even by this very stringent evaluation criteria, the AMC potential performs rather well. It is worth noting that the distance maps look somewhat more native-like than the direct superpositions. The best structures for the test set proteins, as indicated in Table 5, have $Q = 0.35$, $Q = 0.31$, and $Q = 0.28$. It is possible to define a distance map overlap as:

$$(N(N-1))^{-1}\sum_{i\neq j}\theta_{ij},$$

where θ_{ij} is 1 when residues i and j have the same state (contact, no contact) in the native and predicted structure and is 0 otherwise. The corresponding distance map overlaps are 0.77 for 1e4f, 0.76 for 1i74(a), and 0.78 for 1ful(a). It is perhaps unsurprising that the AMC potential would be more successful at predicting the set of pair distances than it is at predicting the global structure. The backbone used in the simulation is highly schematic. Given the success at predicting the inter-residue contacts, it would be interesting to see how much improvement can be achieved with a more elaborate description of the backbone or even the substitution of segments from experimental structures subject to the predicted pair constraints paralleling a fragment assembly method (25).



View larger version (46K):
[in this window]
[in a new window]

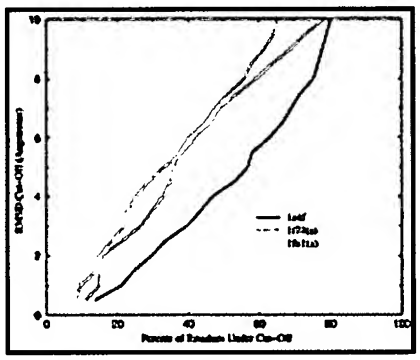
Fig. 1. Superpositions of Q_{best} structures onto the native state. The native C_α trace is shown as lines and the predicted trace as a solid ribbon. (*A*) Protein 1e4f. (*B*) Protein 1i74(a). (*C*) Protein 1ful(a).

The secondary structure bias, in its present form, can sometimes lead to interesting failures. In the case of 1ful(a), the break in the native helix at residue 60 facilitates a turn that the predicted structure lacks. The 1D prediction that enters into the bias has residue 60 as helical. In this case the rather strong bias to the predicted secondary structure that we have used (4ϵ) is a disadvantage. It is obviously possible to choose different, even optimized, weights for this term.

The global superposition of two structures can often fail to highlight significant segments of correct native structure. We have submitted the best Q structures to the **LGA (Local-Global Alignment)** server (http://PredictionCenter.llnl.gov/local/lga), and the results are given in Table 6. The predictions are evaluated according to two measures, LCS and GDT. LCS is the longest continuous (along the sequence) segment that can be superimposed on the native structure without exceeding a rmsd cutoff. The global distance test (GDT) represents the largest number of residues that lie within a distance cutoff of their correct positions. The set of residues need not be contiguous. In all three cases large portions of the prediction are correct to within the cutoff. We have also used CE to align the predicted and native structure. Note that in all three cases the predicted structure is more similar to the native than any of the database structures, thus demonstrating the ability of the potential to generalize from incorrect scaffolds. The scores of local similarity will, of course, depend on the chosen cutoff. Fig. 2 is a Hubbard plot of the percent of residues below the cutoff, as a function of the cutoff distance.

Table 6. Results of **LGA** server analysis of test set predictions

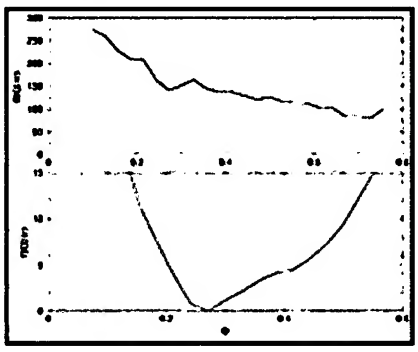
View this table:
[in this window]
[in a new window]



View larger version (12K):
[in this window]
[in a new window]

Fig. 2. GDT as function of cutoff distance.

Although such successful predictions are encouraging, a more complete characterization of the AMC potential requires knowledge of the free energy as a function of similarity to the native state. We can calculate the free energy surface as a function of Q by means of the multiple histogram technique (26). The optimization procedure outlined above is expected to yield a free energy surface that is shaped like a rough funnel toward the native state. In Fig. 3 we show the energy and free energy of 1e4f. The energy declines steadily until relatively high values of Q, indicating that the free energy surface is largely funnel like, with the protein trading energy for entropy as it moves toward the native state. The energy gain is not sufficient to completely balance the loss in entropy, as indicated by the relatively low Q value at the minimum. However, structures with $Q > 0.4$ are certainly accessible within moderate amounts of computation time.

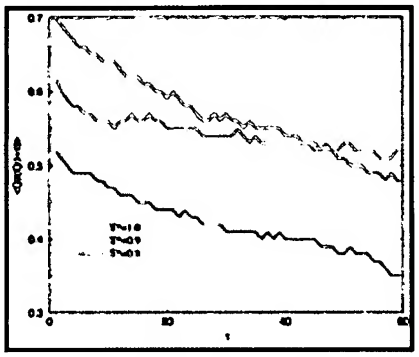


View larger version (9K):
[in this window]
[in a new window]

Fig. 3. Energy and free energy as a function of Q for protein 1e4f (CASP4 Target 0089). $T^* = 1.0$.

For a well-funneled landscape, it is expected that the minima will shift to higher Q as the temperature is lowered. There is a practical problem, however, with simulating at temperatures much lower than those studied here. As the temperature decreases, escape from non-native traps is slowed. At a low enough temperature, we encounter a glass transition, below which the protein is localized to a single basin. Even before that point, however, escape times can become long enough that the finite-time simulations we have performed fall out of equilibrium (27). Fig. 4 shows the Q autocorrelation functions for runs at several different temperatures. It is clear that much below $T^* = 1.0$ the simulation is exploring an increasingly small amount of configuration space. The glass transition therefore limits the simulation to intermediate temperature even though

the minima in free energy would be expected to shift to higher Q as the temperature is lowered. The energy curve shown in Fig. 4 is not monotonically decreasing in Q but is flat above $Q \approx 0.6$. We have previously discussed the practical implications of this caldera-like shape for the sampling of predicted structures (9).



View larger version (9K):
[in this window]
[in a new window]

Fig. 4. Q autocorrelation functions.

Conclusion

We have described a potential energy function for the prediction of α/β protein structures without resorting to information from known, homologous structures. Using ideas from energy landscape theory, we have optimized the parameters of the potential to yield a free energy surface, which is as near to a smooth funnel as is possible given our encoding. The resulting potential performs well in tests on short- to medium-length proteins unrelated to the structures on which it was trained.

Abbreviations

AMC, associative memory and contact; CE, combinatorial extension; rmsd, rms deviation.

Footnotes

¶ To whom correspondence should be addressed. E-mail: pwolynes@ucsd.edu .

Appendix

The α/β training set was selected to represent the various structural classes appearing in the CATH database (28). The 14 training proteins ranged in length from 53 to 138 residues. The training set consisted of proteins 1igd, 2sni(i), 1snb, 3il8, 1ubi, 1pht, 1poh, 1tig, 2acy, 1frd, 1opc, 1rds, 3chy, and 5nul. The scaffolds were a subset of the α/β chains appearing in the Protein Data Bank select 2001 list (29). Structures determined by NMR, those with resolution >3.0 Å, and those with length >200 residues were removed. This process resulted in a list of 168 proteins from which the memory proteins were

selected. The selection process eliminated any memory protein with structural overlap > Q > 0.4 to the training protein to which it was aligned. The final memory set consisted of the top 138 scoring **alignments** to unrelated scaffolds.

▼ References

References

▲ Top
▲ Abstract
▲ Introduction
▲ Materials and Methods
▲ Constrained Self-Consistent...
▲ Results and Discussion
▲ Conclusion
▲ Appendix
▪ References

1. Sanchez, R. , Pieper, U. , Melo, F. , Eswar, N. , Marti-Renom, M. A. , Madhusudhan, M. , Mirkovic, N. & Sali, A. (2000) *Nat. Struct. Biol.* 7, 986-990[CrossRef][Medline] .
2. Hardin, C. , Pogorelov, T. & Luthey-Schulten, Z. (2002) *Curr. Opin. Struct. Biol.* **12**, 176-181[CrossRef][ISI] [Medline] .
3. Anfinsen, C. (1973) *Science* **96**, 223-230.
4. Kim, T. , Simmons, R. , Bonneau, I. R. & Baker, D. (1999) *Proteins Struct. Funct. Genet.* **37, Suppl. 3**, 171-176 [CrossRef][Medline] .
5. Ortiz, A. R. , Kolinski, A. , Rotkiewicz, P. , Ilkowski, B. & Skolnick, J. (1999) *Proteins* **37, Suppl. 3**, 177-185 [CrossRef][Medline] .
6. Pillardy, J. , Czaplewski, C. , Liwo, A. , Lee, J. , Ripoll, D. R. , Kazmierkiewicz, R. , Oldziej, S. , Wedemeyer, W. J. , Gibson, K. D. , Arnautova, Y. A. , *et al.* (2001) *Proc. Natl. Acad. Sci. USA* **98**, 2329-2333[Abstract/Free Full Text].
7. Hardin, C. , Eastwood, M. , Luthey-Schulten, Z. & Wolynes, P. G. (2000) *Proc. Natl. Acad. Sci. USA* **97**, 14235-14240 [Abstract/Free Full Text].
8. Betancourt, M. & Skolnick, J. (2001) *J. Comput. Chem.* **22**, 339-353[CrossRef][ISI].
9. Eastwood, M. P. , Hardin, C. , Luthey-Schulten, Z. & Wolynes, P. G. (2001) *IBM Systems Res.* **45**, 475-497.
10. Onuchic, J. N. , Luthey-Schulten, Z. & Wolynes, P. G. (1997) *Annu. Rev. Phys. Chem.* **48**, 539-594.
11. Bryngelson, J. D. & Wolynes, P. G. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 7524-7528[Abstract/Free Full Text].
12. Friedrichs, M. S. & Wolynes, P. G. (1989) *Science* **246**, 371-373[ISI].
13. Koretke, K. K. , Luthey-Schulten, Z. & Wolynes, P. G. (1998) *Proc. Natl. Acad. Sci. USA* **95**, 2932-2937 [Abstract/Free Full Text].
14. Hardin, C. , Eastwood, M. , Prentiss, M. , Luthey-Schulten, Z. & Wolynes, P. G. (2002) *J. Comput. Chem.* **23**, 138-146 [CrossRef][ISI][Medline] .
15. Koretke, K. K. , Luthey-Schulten, Z. & Wolynes, P. G. (1996) *Protein Sci.* **5**, 1043-1059[Abstract].
16. Eastwood, M. , Hardin, C. , Luthey-Schulten, Z. & Wolynes, P. G. (2002) *J. Chem. Phys.* **117**, 4602-4615[CrossRef] [ISI].
17. Merkel, J. , Sturtevant, J. & Regan, L. (1999) *Struct. Folding Des.* **7**, 1333-1343[ISI].
18. Wouters, M. & Curmi, P. (1995) *Proteins* **22**, 119-131[CrossRef][ISI][Medline] .
19. Cuff, J. & Barton, G. (1999) *Proteins* **34**, 508-519[CrossRef][ISI][Medline] .
20. Reva, B. A. , Skolnick, J. & Finkelstein, A. V. (1999) *Proteins* **35**, 353-359[CrossRef][ISI][Medline] .
21. Altschul, S. , Madden, T. , Schaffer, A. , Zhang, J. , Zhang, Z. , Miller, W. & Lipman, D. (1997) *Nucleic Acids Res.* **25**, 3389-3402[Abstract/Free Full Text].
22. Ryckaert, J. , Ciccotti, G. & Berendsen, H. (1977) *J. Comput. Phys.* **23**, 327-341[CrossRef][ISI].
23. Goldstein, R. A. , Luthey-Schulten, Z. A. & Wolynes, P. G. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 4918-4922 [Abstract/Free Full Text].
24. Shindyalov, I. & Bourne, P. (1998) *Protein Eng.* **11**, 739-747[Abstract/Free Full Text].
25. Simons, K. , Kooperberg, C. , Huang, E. & Baker, D. (1997) *J. Mol. Biol.* **268**, 209-225[CrossRef][ISI][Medline] .

26. Ferrenberg, A. & Swendsen, R. (1989) *Phys. Rev Lett.* **63**, 1195-1198[CrossRef][ISI][Medline] .
27. Plotkin, S. S. , Wang, J. & Wolynes, P. G. (1997) *J. Chem. Phys.* **106**, 2932-2948[CrossRef][ISI].
28. Orengo, C. A. , Michie, A. D. , Jones, S. , Jones, D. T. , Swindells, M. B. & Thronton, J. M. (1997) *Structure (London)* **5**, 1093-1108[Medline] .
29. Hobohm, U. , Scharf, M. , Schneider, R. & Sander, C. (1992) *Protein Sci.* **1**, 409-417[Abstract].

www.pnas.org/cgi/doi/10.1073/pnas.252753899

Originally published In Press as doi:10.1074/jbc.M301549200 on March 10, 2003

J. Biol. Chem., Vol. 278, Issue 20, 18588-18596, May 16, 2003

Crystal Structure of *Mycobacterium tuberculosis* Diaminopimelate Decarboxylase, an Essential Enzyme in Bacterial Lysine Biosynthesis*

Kuppan Gokulan‡, Bernhard Rupp‡§, Martin S. Pavelka Jr.¶‖, William R. Jacobs Jr.¶, and James C. Sacchettini‡‡ ‡**

From the ‡ Department of Biochemistry and Biophysics, Texas A&M University, College Station, Texas 77843-2128, the § Biology and Biotechnology Research Program, L-448, Lawrence Livermore National Laboratory, Livermore, California 94551, and the ¶ Department of Microbiology and Immunology and the ** Howard Hughes Medical Institute, Albert Einstein College of Medicine, The Bronx, New York 10461

Received for publication, February 13, 2003, and in revised form, March 3, 2003

ABSTRACT

The *Mycobacterium tuberculosis lysA* gene encodes the enzyme *meso*-diaminopimelate decarboxylase (DAPDC), a pyridoxal-5'-phosphate (PLP)-dependent enzyme. The enzyme catalyzes the final step in the lysine biosynthetic pathway converting *meso*-diaminopimelic acid (DAP) to L-lysine. The *lysA* gene of *M. tuberculosis* H37Rv has been established as essential for bacterial survival in immunocompromised mice, demonstrating that *de novo* biosynthesis of lysine is essential for *in vivo* viability. Drugs targeted against DAPDC could be efficient anti-tuberculosis drugs, and the three-dimensional structure of DAPDC from *M. tuberculosis* complexed with reaction product lysine and the ternary complex with PLP and lysine in the active site has been determined. The first structure of a DAPDC confirms its classification as a fold type III PLP-dependent enzyme. The structure shows a stable 2-fold dimer in head-to-tail arrangement of a triose-phosphate isomerase (TIM) barrel-like α/β domain and a C-terminal β sheet domain, similar to the ornithine decarboxylase (ODC) fold family. PLP is covalently bound via an internal aldimine, and residues from both domains and both subunits contribute to the binding pocket. Comparison of the structure with eukaryotic ODCs, in particular with a di-fluoromethyl ornithine (DMFO)-bound ODC from *Trypanosoma bruceii*, indicates that corresponding DAP-analogues might be potential inhibitors for mycobacterial DAPDCs.

INTRODUCTION

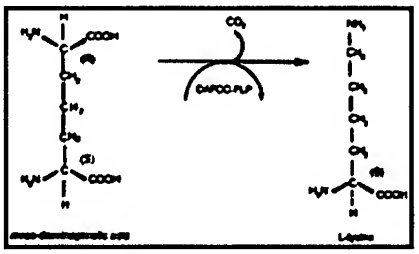
The final step in the bacterial lysine biosynthetic pathway is carried-out by *meso*-DAP[1] decarboxylase (DAPDC), encoded by the *lysA* gene. DAPDC is a vitamin B_6-dependent enzyme that stereospecifically converts *meso*-DAP to L-lysine

▼ RESULTS AND DISCUSSION
▼ REFERENCES

(Scheme 1). Like most enzyme-catalyzed decarboxylation reactions, the conversion of DAP to lysine is not reversible. The enzyme is of interest because of its importance in bacterial growth and survival. Lysine is required in protein biosynthesis and is essential for bacterial viability and development. The lysine precursor DAP itself is used as a structural cross-linking component of the peptidoglycan layer of Gram-negative, Gram-positive (except Gram-positive *cocci*), and mycobacterial cell walls (1). DAP cross-links provide stability to the cell wall and confer resistance to intracellular osmotic pressure (2). DAP can be synthesized by one or more of the following three different pathways: (i) the succinylase pathway, identified in all Gram-negative and Gram-positive bacteria, as well as *Myobacterium tuberculosis*; (ii) the dehydrogenase pathway, utilized by *Bacillus sphaericus, Corynebacterium glutamicum*, and *Brevibacterium* species (3); and (iii) the acetylase pathway, which is limited to certain *Bacillus* species (4). Higher plants also produce lysine using a succinylase pathway (5). The presence of multiple biosynthetic pathways, at least in some bacteria, is probably an indication of the importance of DAP and lysine to bacterial survival. As the substrate and the reaction are not found in mammals, inhibitors of the enzyme may ultimately become leads for therapeutic intervention in bacterial infections (6).

In *Escherichia coli*, the *lysA* gene is transcriptionally controlled by the LysR regulator protein; in the presence of lysine, transcription of the *lysA* gene is repressed (7). In contrast, *M. tuberculosis* does not apparently have a comparable LysR regulator, based on the lack of homologous sequences in the *M. tuberculosis* genomic sequence (8). In *M. tuberculosis*, *C. glutamicum*, and *Brevibacterium lactofermentum*, the *lysA* gene is not in an operon as the second gene in an open reading frame with *argS* (arginyl-tRNA synthetase) (9-12). In *C. glutamicum* the *lysA* gene is constitutively expressed (11), and in the related organism *B. lactofermentum* the *lysA* gene is only weakly suppressed by lysine (12). Based on the evolutionary relationship between these three species of bacteria, we (13) proposed that the expression of the *lysA* gene of *M. tuberculosis* is probably constitutive.

We show in this study that the *lysA* gene is essential for *M. tuberculosis* survival in an immunodeficient SCID (severe combined immunodeficient) mouse model, and we have determined the crystal structure of DAPDC in complex with the coenzyme pyridoxal 5'-phosphate (PLP) and the decarboxylation product lysine as well as DAPDC complexed with only lysine (binary complex). DAPDC is structurally very similar to eukaryotic ornithine decarboxylases (ODCs) (14-16) and, with the exception of a rotation of the C-terminal domain, to *Bacillus stearothermophilus* alanine racemase (AR) (17). Although both DAPDC and ODCs carry-out similar decarboxylation reactions involving pyridoxal-5'-phosphate (PLP) as a cofactor, DAPDC is the only known amino acid decarboxylase that stereospecifically acts on a substrate carbon atom in D-configuration (Scheme 1).



View larger version (12K):
[in this window]
[in a new window]

Scheme 1. Reaction schematic of stereospecific decarboxylation of *meso*-diaminopimelic acid (DAP) to L-lysine via vitamin B_6 (PLP)-dependent DAP-decarboxylase (DAPDC).

▶ EXPERIMENTAL PROCEDURES

▲ TOP
▲ ABSTRACT
▲ INTRODUCTION

Generation and in Vitro Characterization of the lysA Mutant of M. tuberculosis--The *lysA* mutant of *M. tuberculosis*, mc^{2}3026, was previously constructed by

- EXPERIMENTAL PROCEDURES
- RESULTS AND DISCUSSION
- REFERENCES

allelic exchange and has a deletion within the coding region of the *lysA* gene with an inserted γδ resolvase binding site (18). The mutant requires exogenous lysine supplementation at 1 mg/ml and can be complemented to protrophy by a copy of the wild-type *lysA* gene carried on the integrating vector pYUB651. In this work, we performed reversion analysis and were unable to isolate revertants from over 10^{10} *M. tuberculosis* Δ*lysA* cells. This established that the DAPDC activity can not be suppressed by any extragenic mutation and that the viability of the *M. tuberculosis* cells is dependent on this activity.

Clearance of the M. tuberculosis Lysine Auxotroph in SCID Mice-- Female SCID mice were bred at the animal facility of the Albert Einstein College of Medicine. The animals were maintained under barrier conditions and fed sterilized commercial mouse chow and water *ad libitum*. The *M. tuberculosis* strains $mc^2$3026 (Δ*lysA5::res*) and $mc^2$3026 bearing pYUB651 (expressing the wild-type *lysA* gene) (13), were grown in Middlebrook 7H9 broth (Difco) supplemented with 0.05% Tween 80, 0.2% glycerol, and 1× ADS (0.5% bovine serum albumin, fraction V (Roche), 0.2% dextrose, and 0.85% NaCl) or on Middlebrook 7H10 or 7H11 solid medium (Difco) supplemented with 0.2% glycerol and 10% OADC (oleic acid, albumin, dextrose, and catalase; BD Biosciences). Cultures of the lysine auxotroph were supplemented with 1 mg/ml L-lysine (for both liquid and solid media), and 0.05% Tween 80 was added to solid medium. Liquid cultures were grown in 490-cm^2 roller bottles (Corning) at 4-6 rpm. Plates were incubated for 3-6 weeks.

Titered frozen stocks of bacteria were thawed and diluted appropriately in phosphate buffered saline containing 0.05% Tween 80 (PBST). The bacterial suspensions were plated at the time of injection to confirm viable counts. Intravenous injections were given via the tail vein. At various time points post-injection (24 h and once weekly), three mice were sacrificed for each strain, and the lungs, liver, and spleen were removed and homogenized separately in PBST using a Stomacher 80 (Tekmar, Cincinnati, OH). The homogenates were diluted in PBST and plated to determine the number of colony-forming units (CFU)/ml. Note that mice were sacrificed at 24 h post-injection in order to compare the bacterial colony-forming units received by the mice to the colony-forming units in the suspensions at the time of injection. Thus, the bacterial counts reported at time zero represent the viable bacteria present in the mice at 24 h post-injection.

Cloning of the lysA Gene and Expression of M. tuberculosis DAPDC-- A 1.3-kb DNA fragment containing the *lysA* gene (Rv1237, Swiss Prot accession number P31848), was amplified by PCR with *M. tuberculosis* H37Rv genomic DNA as the template, using the following oligonucleotide primers: 5'-AGA GAA GCA TAT GAA CGA GCT GCT GCA CTT AGC GCC GAA TG-3' and 5'-AGA GAA GGC GGC CGC CCT CAC TTC CAA ACT CAG CAA ATC GTC-3'. The amplified DNA fragment was digested with *Nde*I and *Not*I restriction enzymes and subcloned into the corresponding restriction sites in the pET30b vector with a C-terminal His_6 tag. *E. coli* B834 (DE3) Met^- cells were transformed with the lysA-pET30b/His vector. The transformed cells were grown to exponential phase at 37 °C in TB media containing kanamycin. For production of Se-Met labeled protein, the cells were grown in M9 minimal media supplemented with all 19 standard amino acids and selenium-methionine (19). Expression of *lysA* was induced with 1 mM isopropyl-1-thio-β-D-galactopyranoside (IPTG), and cells were harvested after growth for 4-6 h at 16 °C.

DAPDC Purification-- The harvested cells were pelleted and resuspended in buffer A (20 mM Tris-HCl, pH 8.0, and 50 mM imidazole) containing 1 mM phenylmethylsulfonyl fluoride (PMSF) and complete EDTA-free protease inhibitors (Roche Applied Science). The cell mixture was repeatedly sonicated at 4 °C with 30 s pulses, and the cell suspension was centrifuged at 15,000 × *g* for 1 h. The clear supernatant was loaded onto an Amersham Biosciences Hi-trap Ni^{2+} chelating column and washed with 300 ml of buffer A containing 500 mM NaCl. The His-tagged DAPDC was eluted from the nickel affinity column using Buffer B (20 mM Tris-HCl, pH8.0, 500 mM imidazole, and 500 mM NaCl). After purification to near homogeneity by size exclusion chromatography (Amersham Biosciences) on an S-Superdex-200 column, DAPDC was dialyzed against 20 mM Tris buffer (pH 8.0), concentrated to 10 mg/ml, and stored in 20 mM Tris-HCl, pH 8.0, at −80 °C.

Crystallization-- Native and Se-Met-labeled DAPDC (10 mg ml^{-1}) were crystallized at 18 °C by vapor diffusion in hanging drops. Initial crystallization screening was carried out with DAPDC alone, DAPDC incubated with DAP (5 mM) plus PLP (0.2 mM) overnight at 4 °C, and DAPDC plus lysine. Crystallization of DAPDC was only successful in the case of DAPDC supplemented with 5 mM lysine. Crystals (0.2 × 0.3 × 0.3 mm) grew at 18 °C within 3-7 days in 4-μl hanging drops (2 μl of

DAPDC, 10 mg/ml, containing 5 mM of lysine combined with 2 μl of well solution) equilibrated against 500 μl of well solution containing 24% polyethylene glycol mono-methylether 5000 (PEG-MME 5000), 0.1 M MES buffer, pH 6.3, and 60 mM ammonium sulfate. Native DAPDC-lysine crystals were soaked for 3 h in mother liquor containing 0.2 mM PLP to obtain distinctly yellow-colored crystals of the DAPDC-PLP-lysine complex.

Data Collection-- Highly redundant and complete selenium K-edge MAD diffraction data from a single Se-Met-DAPDC/lysine crystal were collected at three wavelengths using an ADSC CCD detector on beamline 14-ID-B at the Advanced Photon Source (APS) of the Argonne National Laboratory (ANL). Crystals mounted in cryo-loops were flash cooled in a N_2 stream (120 K) after brief soaks in 2 μl of mother liquor plus 2 μl of a cryoprotectant composed of 30% dioxane and 20% 2-methyl-2,4-pentanediol (MPD). Native data from DAPDC-PLP-lysine crystals were recorded on APS beamline 19BM using the 3 × 3 segment APS-1 CCD detector. The diffraction data were reduced using DENZO (20), and intensities were scaled with SCALEPACK (20). The reflections were indexed primitive tetragonal (a = b = 111.5 Å, c = 237.7 Å) with Laue symmetry 4/mmm. Examination of the integrated and scaled data indicated tetragonal space group $P4_12_12$ or its enantiomorph $P4_32_12$. Solvent content calculations (21) indicated the presence of either a dimer (V_M, 4.0; V_S, 70%) or a trimer (V_M, 2.8; V_S, 54%) in the asymmetric unit.

Structure Determination-- Experimental phases for DAPDC-lysine were obtained by multiwavelength anomalous diffraction (MAD) phasing (22) (Table I). SHELXD located eight selenium sites in the asymmetric unit consistent with a dimer in the asymmetric unit (23), and SOLVE (24) was used to refine the sites and calculate initial protein phases, resulting in an overall figure of merit of 0.41 for the data in the resolution range of 100-2.8 Å. Further phase improvement with solvent flattening in AUTOSHARP (25) resulted in density-modified maps of high quality showing clear electron density for two molecules of protein in the asymmetric unit. The electron density map was submitted to TEXTAL (26) for automated model building. The TEXTAL model fit 80% of the backbone and 20% of the side chains correctly, with the exception of a stretch of 50 amino acids that were traced in the wrong direction; the remaining backbone model fit well into the electron density of the map. After determining the non-crystallographic symmetry (NCS) operator from the selenium substructure using graphical analysis and refinement with (CCP4) LSQKAB, the electron density was averaged and solvent flattened using DM (27). Starting from the TEXAL tracing, all of the residues of DAPDC except Met-1 could be built into the density-modified and -averaged experimental map using XTALVIEW (28). A final model of high quality was produced after several cycles of manual model building, and NCS restrained maximum likelihood refinement with REFMAC5 (29) against the high remote data set (Table II). A sulfate ion, located at the position of the PLP phosphate moiety, was clearly visible in the electron density. 204 water molecules were manually added during iterative cycles of model building and refinement. Weak electron density for the complexed lysine was visible in each binding pocket of the dimer but was not refined in the Se-Met model.

Table I

View this table: [in this window] [in a new window]

Anomalous data collection and phasing statistics for binary DAPDC-Lys complex

Table II

View this table: [in this window] [in a new window]

Data collection, refinement, and geometry statistics for binary and ternary DAPDC complexes

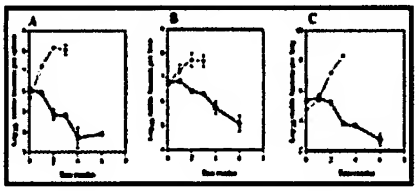
The structure of native DAPDC complexed with PLP and lysine was solved by molecular replacement with EPMR (30) (correlation coefficient 0.60) using the final model of the Se-Met DAPDC-lysine complex as a search model. Bias-minimized electron density maps were obtained using the Shake&wARP (SNW) protocol (31). Clear electron density for both PLP molecules and density for both lysines were visible in the Shake&wARP map prior to any model building. Several cycles of

manual model adjustment and NCS-restrained maximum likelihood refinement in REFMAC5 yielded a final 2.6 Å model of good quality (Table II) for the DAPDC-PLP-lysine complex.

RESULTS AND DISCUSSION

▲ TOP
▲ ABSTRACT
▲ INTRODUCTION
▲ EXPERIMENTAL PROCEDURES
▪ RESULTS AND DISCUSSION
▼ REFERENCES

The lysA Gene Is Required for in Vivo Growth of M. tuberculosis H37Rv-- The lysine auxotrophic strain mc^{2}3026 or the complemented mutant were each introduced (10^6 cells per mouse) into 24 SCID mice by tail vein injections, and groups of three mice each were sacrificed at 1 day post-injection and weekly thereafter until week 6. At each sacrifice, the number of viable bacteria was determined in the spleens, livers, and lungs of the mice. The lysine auxotrophic mutant was cleared from or did not grow in the examined organs of the SCID mice, whereas the complemented strain, mc^{2}3026/pYUB651, multiplied extensively (Fig. 1). In both the spleen and the lung, the number of viable bacteria decreased by three orders of magnitude in 6 weeks (Fig. 1*B*), whereas the decrease of the number of viable bacteria in the lung was only one order of magnitude (Fig. 1*C*). The mice given the complemented *M. tuberculosis* mutant died within 3 weeks, whereas the mice receiving the auxotrophic *M. tuberculosis* mutant did not display any gross organ pathology and survived for the duration of the experiment. Control experiments have demonstrated that immunocompetent C57BL/6 mice can clear an infection with the *M. tuberculosis* lysine auxotroph with the same kinetics as those seen for the clearance of the mutant in the spleen and lungs of the SCID mice (data not shown).



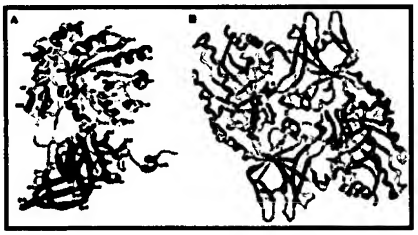
View larger version (12K):
[in this window]
[in a new window]

Fig. 1. Clearance of the lysine auxotrophs in SCID mice. The viable bacterial counts in CFU/ml are shown for the spleens, livers, and lungs of SCID mice injected intravenously with the various mycobacterial strains. Three mice were assayed at each time point. The *error bars* indicate the means ± S.D. Note that the counts at time zero are the counts obtained at 24 h post-injection as described under "Results and Discussion." *Panels A*, *B*, and *C* show the CFU/ml in each organ after injection with 1×10^7 CFU of the Lys$^-$ *M. tuberculosis* mutant mc^{2}3026 (*open squares*), or 1×10^7 CFU of the complemented Lys$^+$ *M. tuberculosis* strain mc^{2}3026/pYUB651 (*closed squares*).

In addition, we tested the frequency of reversion of the *lysA* mutations by growing the mutant in the presence of lysine to mid-log phase of growth, centrifuging it, and resuspending it in media without lysine. The plating of two independent cultures and plating over 10^{10} cells from both cultures yielded no viable colonies, thus establishing that the lysA deletion mutant does not revert and cannot be suppressed by an extragenic mutation. The combination of the *in vitro* and *in vivo* data establishes that DAPDC activity is essential for the viability of *M. tuberculosis* and that *M. tuberculosis* cannot sequester lysine from a mammalian host. We thus reasoned that drugs targeted against DAPDC could be effective anti-tuberculosis agents and pursued the determination of the three-dimensional structure of *M. tuberculosis* DAPDC.

Overview of the M. tuberculosis DAPDC Structure-- The crystal structure of M. tuberculosis DAPDC confirms its classification as a fold type III B_6 dependent enzyme (32). DAPDC has a fold similar to eukaryotic ODCs (14-16), and DAPDC also forms a stable head-to-tail homodimer of practically identical subunits with a coordinate deviation comparable with the overall r.m.s.d. coordinate error for the structure models (0.33 and 0.42 Å, respectively).

Each of the DAPDC subunits (related by proper 2-fold rotation) consists of two ODC-like domains (Fig. 2). Domain I is composed of residues 48-308 forming a α/β barrel comprised of β-strands (β4-β13) and helices (α2-α10). The first 47 residues are located in domain II and contain strands β1, β2, β3, and helix α1, leading into helix α2 of the barrel. The C-terminal domain II contains residues 2-47 (β1, β2, β3, and α1) and 309-446 (α11-α13, strands β14-β21) and forms a mixed β-sheet flanked by α helices. The two structural domains are connected by helix α2 and β13. All of the loops connecting the β strands and α helices were clearly visible in the electron density.

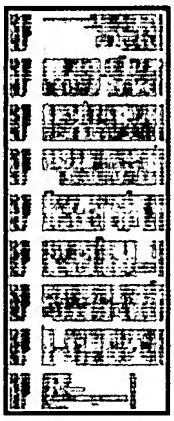


View larger version (40K):
[in this window]
[in a new window]

Fig. 2. Overview of the *M. tuberculosis* DAPDC structure. *A*, ribbon presentation of secondary structure elements. The α/β barrel domain (I) formed of residues 48-308 is shown in *yellow*; the C-terminal domain (II) contains residues 1-47 of the amino-terminal and 309-446 from the C-terminal region and is colored *magenta*. The fold is similar to that of eukaryotic ODCs and classifies *M. tuberculosis* DAPDC as a fold type III B_6-dependent enzyme. *B*, two molecules of DAPDC, related by 2-fold non-crystallographic symmetry, form a stable dimer. Subunit one, same color scheme as in *panel A*. Subunit two is colored *cyan* (N-terminal α/β domain) and *red* (C-terminal domain). Shown in *stick* representation, PLP and lysine, located in the binding pocket formed by dimer interfaces between N-terminal and C-terminal domains. Also shown are the disulfide links between the subunits.

Two identical binding sites are formed by residues of both polypeptide chains of the dimer. The active site is at the interface between the α/β barrel domain of one subunit and the β sheet domain of both subunits. Residues from the α/β barrel are mainly involved in binding PLP, whereas residues from the β sheet domain primarily contribute to substrate binding. Large conformational changes between the binary DAPDC-lysine and ternary DAPDC-PLP-lysine complex are absent (overall C_α coordinate r.m.s.d. 0.42 Å). The only significant differences between the DAPDC complex structures appear near the substrate and cofactor binding sites, discussed below.

Comparison of M. tuberculosis DAPDC with Eucaryotic Ornithine Decarboxylases and Alanine Racemase-- A search for structural alignment using DALI (33) revealed high similarity (Z-value 34.4) with eukaryotic ODCs, enzymes found in the polyamine biosynthetic pathway catalyzing the decarboxylation of ornithine to putrescine and a lower level of structural similarity with AR from *B. stearothermophilus* (Z-value, 18.3). Multiple sequence alignments of known mycobacterial DAPDC sequences, eukaryotic ODCs with known structures, and *B. stearothermopilus* AR are presented in Fig. 3 and summarized together with structural data in Table III. Despite the relatively low level of amino acid sequence identity between eukaryotic ODCs and *M. tuberculosis* DAPDC (~18%), least squares superposition of the structures indicates close resemblance (r.m.s.d. values, ~2.2 Å). Even AR, which shares only 5% identity with DAPDC, superimposes with 2.7 Å r.m.s.d. (Fig. 4). The higher deviation can be attributed largely to a distinct rotation of the AR β-domain respective to the well superimposing α/β barrels (~30°, see also Grishin *et al.* 1999 (34)).



View larger version (53K):
[in this window]
[in a new window]

Fig. 3. Multiple sequence alignment of PLP-dependent enzymes. *Top line* indicates regions of partially conserved or important binding motives or residues. Alignment carried out with ClustalW 1.8.2 (40). Color key: *green*, polar residues; *red*, hydrophobic residues; *blue*, negatively charged; and *magenta*, positively charged.

Table III

View this table: Sequence and structure alignment summary for fold type III PLP dependent enzymes
[in this window]
[in a new window]

View larger version (49K):
[in this window]
[in a new window]

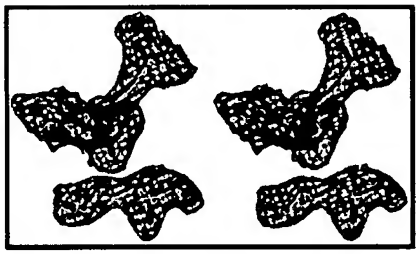
Fig. 4. Backbone superposition of known fold type III PLP-dependent enzyme structures. *Panel A*, color key: *cyan*, *M. tuberculosis* DAPDC; *magenta*, human ODC; *green*, mouse ODC; and *yellow*, *T. brucei*. *Panel B*, superposition of *M. tuberculosis* DAPDC (*cyan*) with *B. stearothermopilus* AR (*red*). The rotation of the AR β-domain relative to the other structures is clearly visible. The superpositions were carried by the Local-Global-Alignment server (Adam Zemla, predictioncenter.llnl.gov/local/lga/lga.html); corresponding r.m.s.d. values are listed in Table III. The figure was prepared using Swiss Pdb Viewer (41) and PovRay (www.povray.org).

The sequence alignments (Fig. 3) also show a decreasing conservation of PLP binding motives from the mycobacterial DAPDCs to the eukaryotic ODCs and AR. The KAFL motif, containing the lysine residue that covalently binds to PLP via Schiff base (internal aldimine) formation, is conserved in procaryotic DAPDCs and eukaryotic ODCs, as is the glycine rich motif (GGG) shown to interact with the phosphate group of PLP. Other conserved motifs include the HIGS motif (thought to be involved in protonation/deprotonation reactions), and the EPGR and CESGD motifs, which are part of the substrate binding regions (35). These motifs, however, with the exception of EPGR, appear not conserved in the structurally related alanine racemase, consistent with its very low sequence identity to the DAPDCs.

Comparison of the large, buried, solvent-accessible surface area at the dimer interface (Table III) indicates that DAPDC (3462 $Å^2$, or 21%) forms the most stable dimer among the fold type III members of known structure. The extensive number of conserved intermolecular contacts and the absence of extended crystal packing contacts (largest contact area between symmetry related subunits in the crystal lattice is 64 $Å^2$) indicate that DAPDC is an obligate dimer. Additional structural support for the dimer as the functional unit comes from the unexpected finding of a disulfide bridge between Cys-93 of one subunit of the dimer and Cys-375 of the other subunit. Intersubunit disulfide bridges are very rare in cytoplasmic proteins, especially in prokaryotes. Cys-93 is found only in mycobacterial DAPDCs but is absent in all other bacterial DAPDCs. Cys-375 also forms a hydrogen bond via its backbone oxygen to the PLP OP3 hydroxyl group of the other subunit and is conserved in all bacterial DAPDCs as well as in other type III B_6-dependent enzymes. Chromatographic experiments further provide chemical evidence that *M. tuberculosis* DAPDC is indeed a stable dimer. DAPDC migrated with an apparent molecular weight consistent with a dimer in gel filtration chromatography experiments, and the disulfide bridge adjoining the two subunits was confirmed by non-reducing SDS-PAGE (not shown). Interestingly, early ultracentrifugation studies reported that *E. coli* DAPDC was a tetramer (36), whereas gel filtration analysis suggested that the *E. coli* DAPDC enzyme was monomeric (37). For *M. tuberculosis* DAPDC, neither the crystal structure nor size exclusion chromatography nor the native SDS gels described above support a monomeric state or the formation of a tetramer.

The PLP-binding Site-- The active site of *M. tuberculosis* DAPDC is located in a shallow, highly hydrophilic cavity between the dimer interfaces with the deep PLP binding pocket located near the C-terminal ends of the β strands of the α/β barrel, similar to other ODCs (14-16). Clear electron density for PLP was visible in the SNW omit maps of the ternary complex and indicated the presence of a covalent C=N link between Lys-72 $N_ε$ and C4A of PLP (Fig. 5).

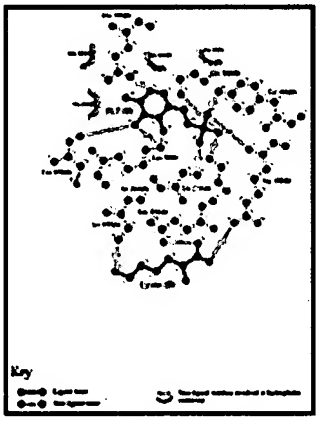
Fig. 5. Electron density in the DAPDC binding cleft for covalently bound PLP and lysine. Both PLP and lysine were omitted from the model before map generation (Shake&wARP map (31) contoured at 1



View larger version (53K):
[in this window]
[in a new window]

σ level). The blob feature in XtalView has been used to limit the display of the electron density within 2 Å of the model. This figure was created by XtalView (28) and rendered with Raster3d (42).

Hydrogen bonds and nonbonding contacts between PLP and DAPDC are summarized in Fig. 6. The oxygen atoms of the PLP phosphate group hydrogen bond with the peptide backbone nitrogen atoms of Gly-258 in the glycine rich motif and those of Gly-302 and Arg-303. OP1 also forms a hydrogen bond with the hydroxyl group of Tyr-405. In the DAPDC-lysine binary complex, a sulfate ion occupies the same position as the phosphate group of PLP in the ternary DAPDC-PLP-lysine structure.



View larger version (22K):
[in this window]
[in a new window]

Fig. 6. Schematic representation of ligand binding interactions in active site pocket of DAPDC. Residues of both homodimer subunits contribute to PLP and to lysine binding. This figure was created by LIGPLOT (43).

In addition to the covalent link to Lys-72 N_ζ, the pyridyl moiety of PLP is positioned by a hydrogen bond to the side-chain carboxylate of Glu-300, which participates in an extended hydrogen bond network with Asp-91 and the conserved residues Asp-254 and His-211. Two histidine residues (114 and 213) and Ala-70 form hydrophobic contacts, with His-213 π-stacking against the *si* face of the pyridine ring. His-114 and Asp-91 are positioned toward the *re* face of the pyridine ring, and both are within hydrogen bonding distance of the carboxylate of Glu-300. The network of interactions around Glu-300 in the binding pocket essentially fixes the position of the imidazole side chains of His-114 and His-211, as well as the carboxylates of Asp-91 and Asp-254 with respect to the pyridine ring of PLP. An additional hydrogen bond to the other subunit of the dimer exists between the O3 hydroxyl group of PLP(B) and the backbone oxygen of the disulfide forming cysteine Cys-375A (Fig. 6).

The side chain of His-213 π-stacks with the *si* side of the pyridyl ring. This residue is conserved in the eukaryotic ODCs; however, His-211 and His-114 are absent and are replaced by serine and alanine or glycine, respectively. In *B. stearothermopilus* AR, the π-stacked His-213 is again conserved via AR His-166, and His-211 and His-114 are replaced by Tyr-164 and Leu-85. The highly variable environment on the *re* face of the pyridyl ring caused by these residue substitutions could play a significant role in fine-tuning the (stereo)specificity and/or pH optimum of the different PLP-

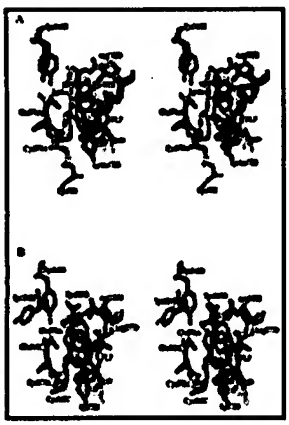
mediated reactions in these enzymes.

Lysine Binding to M. tuberculosis DAPDC-- In the DAPDC-PLP-lysine complex, the density for reaction product lysine could be located in each binding site. In binding site B, the density is very clear and allowed unambiguous positioning and refinement of the lysine molecule (Fig. 5). In site A, the lysine is again oriented similarly to the first site, but its exact position along the channel opening in the binding site is not as clear as for site B. Both lysines are positioned with the side chain toward the *si* face of the PLP pyridyl ring, consistent with decarboxylation occurring on this side of the ring. Residues of domain II of the other subunit (Ser-377A, Glu 376A) participate in lysine binding consistent with the important role of ODC Asp-361 (corresponding to DAPDC Glu-376) that has been demonstrated in Ala mutation studies (38), which show a 2000-fold decrease in substrate binding affinity in mODC. The carboxyl group of lysine is further fixed by conserved residue Arg-303, which participates in PLP binding via backbone N contacts as well. As clearly visible in the electron density Fig. 5, the ε-amino group and CE of lysine are positioned reasonably close (~4.0 Å) to the catalytic Schiff base formed by the Lys-PLP internal aldimine. (Fig. 5). A model of the substrate DAP based on the bound lysine would thus have its (D)-aminoacyl group in a position to interact with the internal aldimine from the *si* side of the pyridoxyl ring as well as with conserved His-213, Arg-161, and possibly Ser-377.

Given the limited 2.6 Å resolution of the present structure, further discussion of the details of the stereospecificity of the decarboxylation mechanism in DAPDC must remain speculative. The structural similarity between the DAPDC binding site and that of eukaryotic ODCs suggests a related mechanism. The catalytic mechanism of the decarboxylation reaction preformed by ODCs has been extensively studied (14, 38). The major difference is that in ODCs the amino acid substrate ornithine is in an L configuration, but DAPDC decarboxylates the D-aminocarboxyl group of *meso*-DAP. Details in the orientation of the D-aminocarboxyl group with respect to the conjugated pyridyl ring system acting as an electron sink as well as stereospecificity of the anchoring of the non-reacting L-aminocarboxyl group through the domain II residues are likely responsible for achieving stereospecific decarboxylation of DAP. Amino acid decarboxylation reactions of fold type III PLP-dependent enzymes generally occur on the *si* side of the pyridyl ring plane (as discussed by Kern *et al.* in 1999 (15)), and evidence exists that the reaction may involve an inversion of the reactive C_α of the substrate (39).

Structural Basis of DAPDC as a Potential Anti-tuberculosis Drug Target-- The comparison of DAPDC with the inhibitor- and product-bound ODC structures (14, 32) of the parasitic flagellate *Trypanosoma brucei* indicates that DAPDC, given that it is essential for *M. tuberculosis* viability, could be a potential anti-mycobacterial drug target. Although there are currently no known drugs that target DAPDC, one of the most widely used drugs used to treat African sleeping sickness is α-difluoromethylornithine (DFMO), a suicide inhibitor that targets *T. brucei* ODC (32). In the crystal structure, DFMO forms the external aldimine linkage with PLP as seen in the product-bound structure (14), but in addition it is covalently bound to the side chain of Cys-360, thus irreversibly blocking the binding site (Fig. 5). A slight backbone torsion, combined with an ~160° rotation of the equivalent Cys-375 SG, suffices to bring DAPDC into practically the same conformation as the DFMO-bound *T. brucei* ODC (Fig. 7) but necessitates the breakage of the intersubunit disulfide bond in DAPDC. It has been proposed that in ODCs DFMO decarboxylation via the internal PLP aldimine followed by elimination of a F^- anion might form a highly reactive electrophilic imine, attacking the nucleophilic Cys-360 thiol group (35). To what degree a reactive imine of a fluorinated DAP analogue might be capable of attacking the Cys-375-Cys-93 disulfide bond, is unknown. It certainly would require a transient conformational rearrangement, probably associated with a slight rotation of PLP, which now has lost its covalent link, to position the reactive imine so that a reaction can take place. Provided the disulfide bond gets broken, the product conformation would closely resemble the arrangement found in DFMO-bound ODC. An energy-minimized model, starting from a DAP molecule placed just as the bound DFMO in the *T. brucei* x-ray structure, shows that the same conformation is conceivable for a putative DAPDC-inhibitor complex, with quite satisfying geometry (Fig. 7). Stereospecificity of the decarboxylation reaction preceding the attack of the reactive imine intermediate would likely require that a DAP analog be stereospecifically fluorinated at the D-aminocarboxyl group of DAP.

Fig. 7. Stereoviews of superpositions of the active sites of the two models. *A*, superposition of energy-minimized models of putative DAPDC-inhibitor (DFDAP) complex (*green* carbon backbone) with ternary DAPDC-PLP-Lysine complex (*cyan* carbon backbone) are shown in stereo. The aminocarboxyl group on the PLP-bound



View larger version (41K):
[in this window]
[in a new window]

DFDAP molecule occupies a position similar to lysine in the ternary DAPDC complex. The DAPDC-DFDAP was modeled covalently bound to Cys-375A, causing speculation that breakage of the Cys-375A to Cys-93B intersubunit disulfide bond could occur through an attack of a highly reactive fluorinated imine intermediate (35). *B*, superposition of the putative DAPDC-inhibitor (DFDAP) complex (*green* carbon backbone) with the *T. bruceii* ODC-DFMO complex (*cyan* carbon backbone), showing the similarity in the overall geometry of the bound inhibitors in stereo. *T. bruceii* ODC-DFMO was superimposed onto the structure of DAPDC to achieve a crude positioning of the PLP-DFMO complex in the active site of the tuberculosis (TB) enzyme. The PLP-DFMO complex was extended to the corresponding bound PLP-DFDAP analog, and the starting position was adjusted. Hydrogens were added, and the docked model was refined further with BioMedCaChe (v.6.0a1). Valence and hybridization checks were enabled and improved hydrogen bond lengths and van der Waals interactions. The structure of DAPDC with the bound PLP-DFDAP analog complex was optimized using the Bio-MM2 molecular mechanics engine in CaChe.

ACKNOWLEDGEMENTS

We acknowledge Katherine Kantardjieff, Center for Molecular Structure, California State University, Fullerton, for ligand docking and minimization calculations on the putative DAPDC-inhibitor complex. B. R. thanks James C. Sacchettini, Texas A&M University, and the Lawrence Livermore National Laboratory for support of his sabbatical leave at Texas A&M University. We thank Ms. Linda Fisher for preparation of the manuscript and Ms. Stephanie Swanson for technical assistance.

FOOTNOTES

* This work was funded by National Institutes of Health Grant P50 GM62410 (Tuberculosis Structural Genomics) and the Robert A. Welch Foundation. Use of the Argonne National Laboratory Structural Biology Center beamlines at the Advanced Photon Source was supported by the United States Department of Energy Office of Energy Research under contract number W-31-109-ENG-38. Use of BioCARS Sector 14 was supported by the National Institutes of Health, National Center for Research Resources. The Lawrence Livermore National Laboratory is operated by University of California for the United States Department of Energy under contract W-7405-ENG-48.The costs of publication of this article were defrayed in part by the payment of page charges. The article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

‖ Present address: Dept. of Microbiology and Immunology, University of Rochester Medical Center, Rochester, NY 14642. Supported by a Burroughs Wellcome Fund Career Award in the Biomedical Sciences.

To whom correspondence should be addressed. Tel.:979-862-7636; Fax: 979-862-7638; E-mail:sacchett@tamu.edu.

Published, JBC Papers in Press, March 10, 2003, DOI 10.1074/jbc.M301549200

ABBREVIATIONS

The abbreviations used are: DAP, *meso*-diaminopimelic acid; DAPDC, *meso*-diaminopimelic acid decarboxylase; PLP, pyridoxal 5'-phosphate; ODC ornithine decarboxylase, SCID, severe combined immunodeficient; PBST, phosphate-buffered saline with Tween 80; CFU, colony-forming unit; MES, 4-morpholineethanesulfonic acid; APS, Advanced Proton Source; r.m.s.d., root mean square deviation; AR, alanine racemase; DFMO, α-difluoromethylornithine.

REFERENCES

▲ TOP
▲ ABSTRACT
▲ INTRODUCTION
▲ EXPERIMENTAL PROCEDURES
▲ RESULTS AND DISCUSSION
▪ REFERENCES

1. Cummins, C. S., and Harris, H. (1996) *J. Gen. Microbiol.* **14**, 583-600
2. Strominger, J. L. (1962) *Fed. Proc.* **21**, 134-143[Medline] [Order article via Infotrieve]
3. Misono, H., Nagasaki, S., and Soda, K. (1986) *Agric. Biol. Chem.* **50**, 1455-1460
4. Sundharadas, G., and Gilvarg, C. (1967) *J. Biol. Chem.* **242**, 3983-3984[Abstract/Free Full Text]
5. Chatterjee, S. P., Singh, B. K., and Gilvarg, C. (1994) *Plant Mol. Biol.* **26**, 285-290[CrossRef][Medline] [Order article via Infotrieve]
6. McCann, P. P., and Pegg, A. E. (1992) *Pharmacol. Ther.* **54**, 195-215[CrossRef][Medline] [Order article via Infotrieve]
7. Stagier, P., Brone, F., Richard, F., Richard, C., and Patte, J. C. (1983) *J. Bacteriol.* **156**, 1198-1203 [Abstract/Free Full Text]
8. Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S. V., Eiglmeier, K., Gas, S., Barry, C. E., III, Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T., Gentles, S., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Krogh, A., McLean, J., Moule, S., Murphy, L., Oliver, K., Osborne, J., Quail, M. A., Rajandream, M.-A., Rogers, J., Rutter, S., Seeger, K., Skelton, J., Squares, R., Squares, S., Sulston, J. E., Taylor, K., Whitehead, S., and Barrell, B. G. (1998) *Nature* **393**, 537-544[CrossRef] [Medline] [Order article via Infotrieve]
9. Andersen, A. B., and Hansen, E. B. (1993) *Gene* **124**, 105-109[CrossRef][Medline] [Order article via Infotrieve]
10. Sharp, P. M., and Mitchell, K. J. (1993) *Mol. Microbiol.* **8**, 200[CrossRef][Medline] [Order article via Infotrieve]
11. Marcel, T., Archer, J. A., Mengin-Lecreulx, D., and Sinskey, A. J. (1990) *Mol. Microbiol.* **4**, 1819-1830[CrossRef] [Medline] [Order article via Infotrieve]
12. Oguiza, J. A., Malumbres, M., Eriani, G., Pisabarro, A., Mateos, L. M., Martin, F., and Martin, J. F. (1993) *J. Bacteriol.* **175**, 7356-7362[Abstract/Free Full Text]
13. Pavelka, M. S., Jr., and Jacobs, W. R., Jr. (1996) *J. Bacteriol.* **178**, 6496-6507[Abstract/Free Full Text]
14. Jackson, L. K., Brooks, H. B., Osterman, A. L., Goldsmith, E. J., and Phillips, M. A. (2000) *Biochemistry* **39**, 11247-11257[CrossRef][Medline] [Order article via Infotrieve]
15. Kern, A. D., Oliveira, M. A., Coffino, P., and Hackert, M. L. (1999) *Structure Fold. Des.* **7**, 567-581[CrossRef] [Medline] [Order article via Infotrieve]
16. Almrud, J. J., Oliveira, M. A., Kern, A. D., Grishin, N. V., Phillips, M. A., and Hackert, M. L. (2000) *J. Mol. Biol.* **295**, 7-16[CrossRef][Medline] [Order article via Infotrieve]
17. Shaw, J. P., Petsko, G. A., and Ringe, D. (1997) *Biochemistry* **36**, 1329-1342[CrossRef][Medline] [Order article via Infotrieve]
18. Pavelka, M. S., Jr., and Jacobs, W. R., Jr. (1999) *J. Bacteriol.* **181**, 4780-4789[Abstract/Free Full Text]
19. Miller, J. H. (1972) *Experiments in Molecular Genetics* , pp. 431-435, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY
20. Otwinowski, Z., and Minor, W. (1997) *Methods Enzymol.* **276**, 307-326
21. Matthews, B. W. (1968) *J. Mol. Biol.* **33**, 491-497[CrossRef][Medline] [Order article via Infotrieve]
22. Hendickson, W. A., and Ogata, C. M. (1997) *Methods Enzymol.* **276**, 494-523
23. Sheldrick, G. M., and Gould, R. O. (1995) *Acta Crystallogr. Sect. B Struct. Crystallogr. Cryst. Chem.* **51**, 423-431 [CrossRef]
24. Terwilliger, T. C., and Berendzen, J. (1999) *Acta Crystallogr. Sect. D Biol. Crystallogr.* **55**, 849-861[CrossRef] [Medline] [Order article via Infotrieve]
25. Cowtan, K. D., and Main, P. (1996) *Acta Crystallogr. Sect. D Biol. Crystallogr.* **52**, 43-48[CrossRef][Medline] [Order article via Infotrieve]
26. Ioerger, T. R., Holton, T., Christopher, J. A., and Sacchettini, J. C. (1999) *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 130137
27. Cowtan, K. D., and Zhang, K. Y. J. (1999) *Prog. Biophys. Mol. Biol.* **72**, 245-270[CrossRef][Medline] [Order article via Infotrieve]

28. McRee, D. E. (1999) *J. Struct. Biol.* **125**, 156-165[CrossRef][Medline] [Order article via Infotrieve]
29. Murshudov, G. N., Vagin, A. A., and Dodson, E. J. (1997) *Acta Crystallogr. Sect. D Biol. Crystallogr.* **53**, 240-255 [CrossRef][Medline] [Order article via Infotrieve]
30. Kissinger, C. R., Gehlharr, D. K., and Fogel, D. B. (1999) *Acta Crystallogr. Sect. D Biol. Crystallogr.* **55**, 484-491 [CrossRef][Medline] [Order article via Infotrieve]
31. Kantardjieff, K. A., Höchtl, P., Segelke, B. W., Tao, F.-M., and Rupp, B. (2002) *Acta Crystallogr. Sect. D Biol. Crystallogr.* **58**, 735-743[CrossRef][Medline] [Order article via Infotrieve]
32. Grishin, N. V., Phillips, M. A., and Goldsmith, E. J. (1995) *Protein Sci.* **4**, 1291-1304[Abstract]
33. Holm, L., and Sander, C. (1995) *Trends Biochem. Sci.* **20**, 478-480[CrossRef][Medline] [Order article via Infotrieve]
34. Grishin, N. V., Osterman, A. L., Brooks, H. B., Phillips, M. A., and Goldsmith, E. J. (1999) *Biochemistry* **38**, 15174-15184[CrossRef][Medline] [Order article via Infotrieve]
35. Poulin, R., Lu, L., Ackermann, B., Bey, P., and Pegg, A. E. (1992) *J. Biol. Chem.* **267**, 150-158[Abstract/Free Full Text]
36. White, P., and Kelley, B. (1965) *Biochem. J.* **96**, 75-84[Medline] [Order article via Infotrieve]
37. Bourot, S., Sire, O., Trautwetter, A., Touze, T., Wu, L. F., Blanco, C., and Bernard, T. (2000) *J. Biol. Chem.* **275**, 1050-1056[Abstract/Free Full Text]
38. Osterman, A. L., Kinch, L. N., Grishin, N. V., and Phillips, M. A. (1995) *J. Biol. Chem.* **270**, 11797-11802 [Abstract/Free Full Text]
39. Asada, Y., Tanizawa, K., Nakamura, K., Moriguchi, M., and Soda, K. (1984) *J. Biochem. (Tokyo)* **95**, 277-282 [Abstract/Free Full Text]
40. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) *Nucleic Acids Res.* **22**, 4673-4680[Abstract/Free Full Text]
41. Guex, N., and Peitsch, M. C. (1997) *Electrophoresis* **18**, 2714-2723[CrossRef][Medline] [Order article via Infotrieve]
42. Merritt, E. A., and Murphy, M. E. P. (1994) *Acta Crystallogr. Sect. D Biol. Crystallogr.* **50**, 869-873[CrossRef] [Medline] [Order article via Infotrieve]
43. Wallace, A. C., Laskowski, R. A., and Thornton, J. M. (1995) *Protein Eng.* **8**, 127-134[Abstract/Free Full Text]
44. Engh, R., and Huber, R. (1991) *Acta Crystallogr. Sect. A* **47**, 392-400[CrossRef]
45. Cruickshank, D. W. J. (1999) *Acta Crystallogr. Sect. D Biol. Crystallogr.* **55**, 583-601[CrossRef][Medline] [Order article via Infotrieve]

This article has been cited by other articles: (Search Google Scholar for Other Citing Articles)

- Citation Map

Services

- Email this article to a friend
- Similar articles in this journal
- Similar articles in PubMed
- Alert me to new issues of the journal
- Download to citation manager
- © Get Permissions

Google Scholar

- Articles by Gokulan, K.
- Articles by Sacchettini, J. C.
- Articles citing this Article
- Search for Related Content

PubMed

- PubMed Citation
- Articles by Gokulan, K.
- Articles by Sacchettini, J. C.

HOME HELP FEEDBACK SUBSCRIPTIONS ARCHIVE SEARCH TABLE OF CONTENTS

All ASBMB Journals **Molecular and Cellular Proteomics**

Journal of Lipid Research

THE JOURNAL OF BIOLOGICAL CHEMISTRY VOL. 282, NO. 13, pp. 10018–10027, March 30, 2007

Structural and Functional Basis of CXCL12 (Stromal Cell-derived Factor-1α) Binding to Heparin*

Received for publication, September 12, 2006, and in revised form, January 19, 2007 Published, JBC Papers in Press, January 29, 2007, DOI 10.1074/jbc.M608796200

James W. Murphy[‡1], Yoonsang Cho[‡], Aristidis Sachpatzidis[‡], Chengpeng Fan[‡], Michael E. Hodsdon[§], and Elias Lolis[‡2]

From the [‡]Department of Pharmacology, Yale University School of Medicine, New Haven, Connecticut 06520-8066 and the [§]Department of Laboratory Medicine, Yale University School of Medicine, New Haven, Connecticut 06520-8035

CXCL12 (SDF-1α) and CXCR4 are critical for embryonic development and cellular migration in adults. These proteins are involved in HIV-1 infection, cancer metastasis, and WHIM disease. Sequestration and presentation of CXCL12 to CXCR4 by glycosaminoglycans (GAGs) is proposed to be important for receptor activation. Mutagenesis has identified CXCL12 residues that bind to heparin. However, the molecular details of this interaction have not yet been determined. Here we demonstrate that soluble heparin and heparan sulfate negatively affect CXCL12-mediated *in vitro* chemotaxis. We also show that a cluster of basic residues in the dimer interface is required for chemotaxis and is a target for inhibition by heparin. We present structural evidence for binding of an unsaturated heparin disaccharide to CXCL12 attained through solution NMR spectroscopy and x-ray crystallography. Increasing concentrations of the disaccharide altered the two-dimensional ^{1}H-^{15}N-HSQC spectra of CXCL12, which identified two clusters of residues. One cluster corresponds to β-strands in the dimer interface. The second includes the amino-terminal loop and the α-helix. In the x-ray structure two unsaturated disaccharides are present. One is in the dimer interface with direct contacts between residues His^{25}, Lys^{27}, and Arg^{41} of CXCL12 and the heparin disaccharide. The second disaccharide contacts Ala^{20}, Arg^{21}, Asn^{30}, and Lys^{64}. This is the first x-ray structure of a CXC class chemokine in complex with glycosaminoglycans. Based on the observation of two heparin binding sites, we propose a mechanism in which GAGs bind around CXCL12 dimers as they sequester and present CXCL12 to CXCR4.

Chemokines are a superfamily of 8–11-kDa secreted chemotactic cytokines, which are modulated by glycosaminoglycans (GAGs),[3] activate chemokine receptors, and direct cellular migration (1). The 43 known human chemokines are divided into two major and two minor families. All have a three-dimensional structure composed of a three-stranded β-sheet followed by an α-helix, which is stabilized by conserved cysteine residues forming typically two disulfide bonds (Fig. 1*A*) (2). The chemokine CXCL12 (CXCL12) is the sole physiological agonist for CXCR4 (3). CXCL12 has been observed as a dimer in several crystal structures (4, 5). However, it has been reported as both a monomer and as a dimer in solution (6, 7). The oligomeric state is modulated by pH, concentration, and the presence of divalent cations and heparin (8). The CXCL12/CXCR4 signaling axis is critically involved in a wide variety of physiological functions. These include margination of neutrophils into the site of infection (9, 10), embryonic development (11), mobilization and directed migration of stem cells (12, 13), and neurological function (14). CXCR4 is involved in HIV-1 infection and HIV-associated dementia (15, 16), is implicated in cancer metastasis (17–19), and genetic mutants are responsible for WHIM disease (20). There is mounting evidence that glycosaminoglycans modulate the activity of chemokines via direct interactions (21). Although the complete functionality of these interactions is not yet completely understood, it is believed that GAGs and proteoglycans sequester chemokines (22). This sequestering increases dimerization and local concentrations, in turn leading to the formation of a chemokine gradient (21–23). In addition, heparan sulfate presents CXCL12 to CXCR4-expressing leukocytes (22). Another important functional aspect of heparin binding to CXCL12 is protection from proteolysis by dipeptidyl peptidase IV (CD26), which removes the NH_2-terminal two residues thus inactivating CXCL12 (24).

CXCL12 was originally shown to possess affinity for heparin and heparan sulfate by binding to heparin affinity resin (25). It was proposed that positively charged CXCL12 and other chemokines interact with negatively charged GAGs through nonspecific electrostatic interactions. However, this interaction was subsequently shown to be specific (26). CXCL12 bound independently of its receptor to CXCR4-negative cells through interaction with cell surface glycosaminoglycans. Amara *et al.* (26) proceeded to use site-directed mutagenesis of CXCL12 to identify a BBXB motif on the first β-strand (Lys^{24}, His^{25}, and Lys^{27}) that was critical for this *in vitro* interaction. This triple mutant retained its *in vitro* biological activity of stimulating

* This work was supported in part by National Institutes of Health Grant RO1 AI065029 (to E. L.), the Pilot Project Funding from the Yale Cancer Center (to M. E. H.), the Campbell Foundation (to E. L.), and the Patterson Foundation (E. L.). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

The atomic coordinates and structure factors (code 2NWG) have been deposited in the Protein Data Bank, Research Collaboratory for Structural Bioinformatics, Rutgers University, New Brunswick, NJ (http://www.rcsb.org/).

[1] Supported by a Neuropharmacology Training Program by National Institutes of Health Grant T32 NS007136.

[2] To whom correspondence should be addressed: 333 Cedar St., New Haven, CT 06510-8066. Tel.: 203-785-6233; Fax: 203-737-2027; E-mail: elias.lolis@yale.edu.

[3] The abbreviations used are: GAG, glycosaminoglycan; HEK-293, human embryonic kidney 293; HSQC, heteronuclear single quantum coherence.

intracellular calcium mobilization. However, the mutated CXCL12 had significantly decreased affinity for heparin (26). Interestingly, mutants of the CC chemokines CCL2 (MCP-1), CCL4 (MIP-1β), and CCL5 (RANTES) devoid of heparin binding retain *in vitro* chemotactic activity but lose *in vivo* activity (27).

A model of the CXCL12-heparin complex was based on the mutagenesis studies (28). Sadir *et al.* (28) calculated an electrostatic potential map of CXCL12 and docked a heparin tetradecasaccharide into the positively charged groove. Energy minimization produced a structure in which heparin was modeled to bind quite well in the putative binding site between the subunits of the CXCL12 dimer. In the proposed model, heparin interacted primarily with the basic residues Lys^1, Lys^{24}, His^{25}, Lys^{27}, Arg^{41}, and Lys^{43} and secondarily with Asn^{46} and Gln^{48}.

We determined that heparan sulfate as well as high and low molecular weight heparins have a concentration-dependent negative effect on CXCL12 driven chemotaxis of $CXCR4^+$ CEM-CCRF cells. Our mutational data revealed a cluster of negatively charged residues involved in both chemotaxis and heparin binding. To elucidate the interactions between the proposed residues and heparin, we used NMR spectroscopy to monitor backbone chemical shift perturbations of CXCL12 during a titration with a heparin disaccharide. By using the fully sulfated, unsaturated disaccharide (heparin disaccharide I-S) derived from a fully sulfated heparin (Fig. 1*B*), we were able to use the minimum size that contains the basic structural elements required for CXCL12 binding. One chemical difference to be noted between this disaccharide and a typical heparin oligomer is the presence of a double bond between C4 and C5 of the uronic acid, making the disaccharide non-physiological. However, the relative positions of the sulfate moieties should remain unchanged. Therefore we used this disaccharide as a model for CXCL12-heparin interactions. The addition of the disaccharide affected two clusters of amino acids in a dose-dependent manner in NMR spectroscopy experiments. We then used x-ray crystallography to define the exact binding position of the disaccharides. The results from NMR spectroscopy combined with the crystal structure of CXCL12 in complex with the disaccharides define specific molecular interactions between CXCL12 and heparin. Our results not only support previous findings that heparin binds to basic residues in the dimer interface (26, 28) but also identify an additional binding site.

EXPERIMENTAL PROCEDURES

Reagents—Luria-Bertani medium, ampicillin, isopropyl 1-thio-β-D-galactopyranoside, Triton X-100, and dithiothreitol were purchased from American Bioanalytical (Natick, MA). ^{15}N-ammonium chloride (98% ^{15}N), heparin disaccharide I-S (containing an unsaturated hexauronate), oxidized and reduced glutathione, protease inhibitor mixture, ammonium sulfate, Trizma base (Tris base), and deuterium oxide were purchased from Sigma.

CXCL12 Expression, Folding, and Purification—The gene encoding human CXCL12 was cloned into the NdeI and XhoI restriction sites of the pET-22b expression vector (Novagen®). The resulting plasmid was transformed into *Escherichia coli* BL21(DE3). 1.5 liters of Luria-Bertani media containing 100 μg/ml ampicillin were inoculated and grown to A_{600} of 0.6 and then induced with 1.0 mM isopropyl 1-thio-β-D-galactopyranoside. Induced cultures were grown for an additional 4 h at 37 °C and harvested by centrifugation for 10 min at 5000 × *g*. Cells were resuspended in 1× phosphate-buffered saline (pH 7.4) with 1% Triton X-100, lysed using a French Press and centrifuged for 20 min at 30,000 × *g*. CXCL12 was found exclusively in inclusion bodies. The inclusion bodies were washed three times in wash buffer A (100 mM Tris·HCl, pH 8.0, 5 mM EDTA, 5 mM dithiothreitol, 2 M urea, 2% Triton X-100) and once with wash buffer B (100 mM Tris·HCl, pH 8.0, 5 mM EDTA, 5 mM dithiothreitol). Washed inclusion bodies were solubilized in 6 M guanidine HCl and diluted 1:100 into a refolding buffer (100 mM Tris·HCl, pH 8.0, 5 mM EDTA, 0.2 mM oxidized glutathione, 1 mM reduced glutathione) and stirred at 4 °C overnight. Precipitated material was removed by filtration. Refolded protein was bound to a SP-Sepharose column and eluted with a NaCl gradient. Fractions containing CXCL12 were pooled and were further purified by reverse phase-HPLC. The peak containing CXCL12 was concentrated and lyophilized. CXCL12 was resuspended in sterile ddH_2O (pH 7.0) with 0.1 mM NaN_3 and protease inhibitor mixture prior to use. Protein concentration was determined by direct amino acid analysis at the W. M. Keck facility (Yale University).

^{15}N-Labeled CXCL12—Isotopically labeled ^{15}N-CXCL12 was produced similar to native protein with the exception of the substitution of LB medium with M9 minimal media containing 1.0 g/liters of $^{15}NH_4Cl$ (98% ^{15}N).

Expression and Purification of Mutant CXCL12—For the chemotaxis assays, wild-type and mutant CXCL12 were expressed in HEK-293 cells to ensure proper folding using a mammalian expression vector containing a secretion signal sequence and an Fc tag at the carboxyl terminus of CXCL12 (kind gift of Timothy Springer of Harvard University). Mutagenesis of the plasmid was accomplished using the QuikChange and Multi-site QuikChange procedures (Stratagene). The proper sequence of all mutated plasmids was verified by DNA sequencing at the W. M. Keck facility (Yale University). HEK-293 cells were transformed at partial confluence by the calcium phosphate transfection method. After 24 h, the growth media containing secreted CXCL12-Fc was collected and treated with iodoacetamide to prevent aggregation. CXCL12-Fc was purified using protein A-Sepharose affinity chromatography.

Cell Culture and Chemotaxis Assays—The CCRF-CEM acute lymphoblastic leukemia cell line (29) was obtained from the American Type Culture Collection (CCL-119) and maintained as recommended by the ATCC. Chemotaxis assays were performed as previously described (30). Briefly, semi-confluent cells were harvested and resuspended to ~1 × 10^7 cells/ml. 100-μl Aliquots of cells were then placed in the upper chamber of a transwell (Corning-Costar®) with 5 μm pore size. Varying concentrations of CXCL12, heparin, and heparan sulfate are in the lower chamber. Following a 2-h incubation at 37 °C, cells which migrated to the lower chamber were counted with an electronic particle counter (Coulter®).

NMR Spectroscopy—Samples for NMR spectroscopy contained 2 mM ^{15}N-CXCL12, 10 mM HEPES, pH 7.4, and 5% D_2O

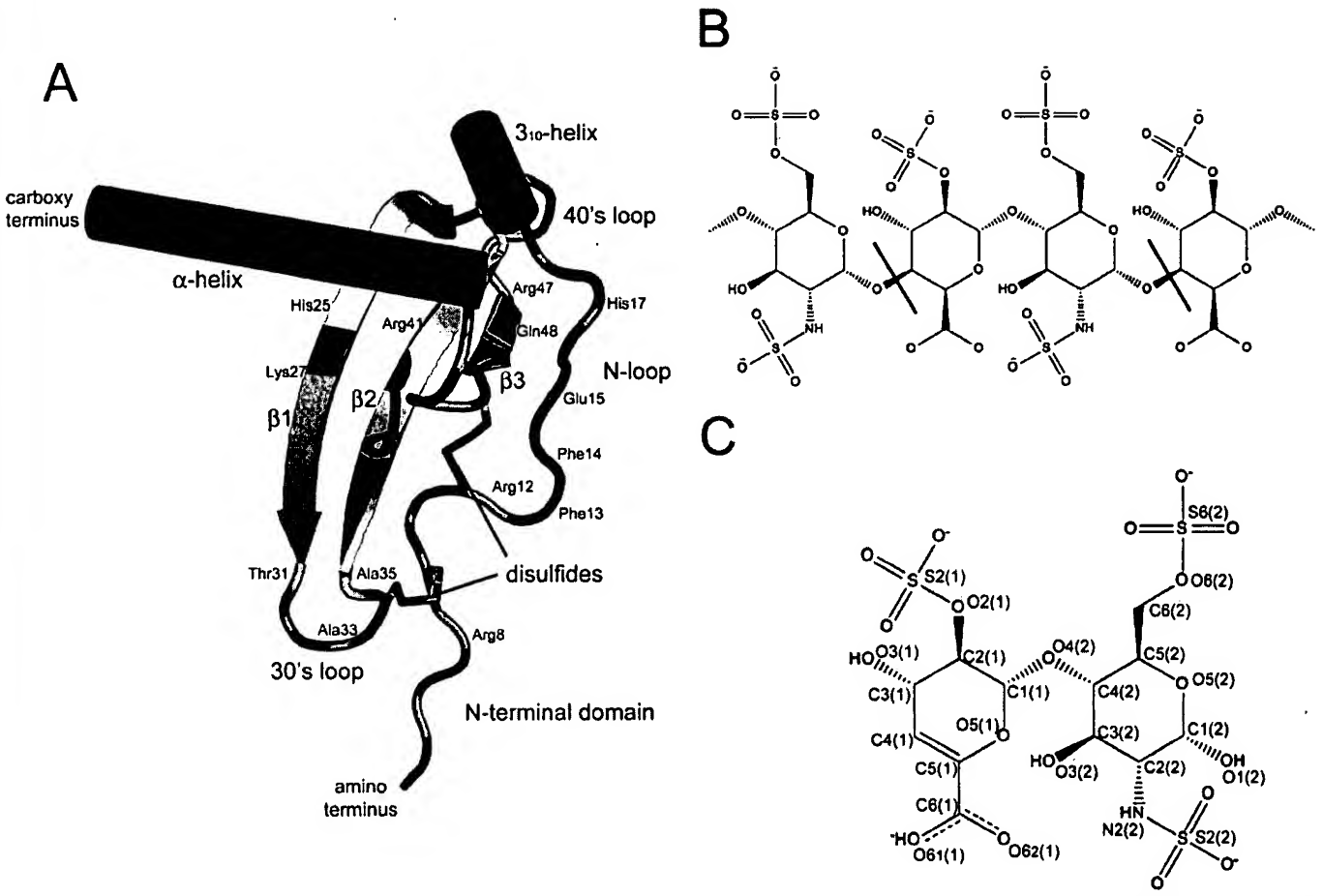


FIGURE 1. *A*, representation of the monomeric form of CXCL12. β-Strands are displayed as *ribbons*, helical regions are displayed as *cylinders* and *loops*, and random coil regions are displayed as a backbone trace. Each structural element is labeled. The two conserved disulfide bonds are displayed in *black*. The position of residues subjected to mutagenesis are indicated by labels. *B*, chemical structure of heparin. The oligomeric chemical structure of a fully sulfated heparin oligomer. A representative tetramer is shown with the polymer continuing through the terminal oxygens. *Dark line segments* indicate which bonds are cleaved by heparinases I and II to form the disaccharide I-S. *C*, heparin disaccharide I-S (α-4-deoxy-L-threo-hex-4-enopyranosyluronic acid-2-sulfate-(1→4)-D-glucosamine-N-sulfo-6-sulfate) as derived from heparin with a sulfate in each of the three R-group positions. Atom numbering follows IUPAC nomenclature for polysaccharide chains. A double bond absent in the heparin oligomer is formed between C^4 and C^5 of the uronic acid creating an unsaturated disaccharide.

with variable concentrations of heparin disaccharide I-S (0–12 mM). Experiments were carried out at 35 °C in a Varian INOVA 600 MHz spectrometer with a 5-mm triple resonance probe equipped with triple-axis (XYZ) pulsed magnetic field gradients. All pulse sequences were taken from the Varian BioPack user library. Spectra were processed and analyzed using the programs nmrPipe (31) and Sparky.[4] Assignment of resonance peaks was done using three-dimensional ^{15}N-TOCSY HSQC and ^{15}N-NOESY HSQC NMR spectra of CXCL12 alone and previously published resonance assignments (33).

Crystallography—CXCL12 was concentrated to 12 mg/ml in water. Protein crystals were grown using the hanging drop method in the previously published condition (2 M ammonium sulfate, 0.1 M Tris·HCl, pH 8.5) at 18 °C (4). The crystals were soaked into 4-μl drops of heparin disaccharide I-S (16 mM) in 20% PEG-8000, 0.1 M Tris·HCl (pH 8.5) and incubated overnight at 18 °C. The crystals were protected in 20% PEG-8000, 0.1 M Tris·HCl (pH 8.5), and 25% glycerol for 1 min and frozen at −180 °C during data collection.

Data Collection and Processing—X-ray diffraction data were collected on an *R*-axis IV detector (Rigaku) at the macromolecular crystallography facility at the Yale University School of Medicine. Data were processed using MOSFLM (34) and SCALA (35). Phaser (36) was used for molecular replacement. The CCP4i software suite (37) and the Crystallography and NMR (CNS) (38) program suite were used for refinement and structure validation. Initial coordinates and structural libraries of the heparin disaccharide were created using the Dundee PRODRG server (39).

RESULTS

Mutagenesis of Wild-type CXCL12-Fc Affects Chemotaxis and Inhibition of Chemotaxis by GAGs—Recombinant, refolded CXCL12 produced in *E. coli* and Fc-tagged CXCL12 produced in HEK-293 cells had equivalent chemotactic activity (data not shown). A series of mutants of CXCL12 were designed to probe the effect of various surface areas in CXCR4 binding and activation (Fig. 1*A*). The EC_{50} values and efficacies of these mutants as compared with the wild-type protein are in Table 1. The chemotactic index is a measure of the chemotactic activity of each protein,

[4] T. D. Goddard and D. G. Kneller, personal communication.

defined as the ratio of the number of cells migrated for each protein concentration to the number of cells migrated in response to assay media without chemokine. Efficacy (%) as used in Table 1 is a measure of the maximum activity of each mutant relative to the maximum chemotaxis of the wild type CXCL12.

The RFFESH motif (residues 12–17) was shown by others to have a role in receptor binding and activation through the use of chimeric chemokines with CXCL12 sequences, receptor binding assays, and intracellular calcium release upon agonist activation (6). We further explored this region with the single mutant R12Q and two double mutants, F13A/F14A and E15Q/H17N. Our results show increases in EC_{50} values ranging from 1.6 to 4.4-fold for the three mutants relative to wild-type CXCL12. Although our results cannot be directly compared with those of Crump *et al.* (6) due to differences in assays (chemotaxis *versus* Ca^{2+} mobilization), we agree that this region is important for CXCL12 function, with no particular residue making a large contribution relative to others. We were surprised, however, by the effect of Arg^8 on the EC_{50}. The maximum activity of three mutants R8Q, R12Q, and R8Q/R12Q is similar to wild type, but these mutants have correspondingly 60, 4, and 120 times higher EC_{50} values than the wild type. The positive charge of Arg^8 is critical for activating CXCR4-mediated chemotaxis. A series of other mutants were designed to address the role of positively charged surface areas identified originally in the CXCL12 crystal structure (4). One of these, a quintuple mutant H25N/K27Q/R41Q/R47Q/Q48N that removed the charges showed a 35% reduction in efficacy with respect to wild type but a similar EC_{50} suggesting that the loss of a large surface area of positive potential has a moderate effect on receptor activation. Alternatively, because this mutant showed a clear resistance to heparin inhibition (see below), its reduced activity may be due to a lack of interaction with native GAGs essential for a maximum chemotactic effect. A triple mutant comprising residues T31V/A33G/A35G showed a small reduction of EC_{50} and efficacy suggesting a possible role of these residues in CXCL12-CXCR4 interaction.

TABLE 1
Effect of CXCL12 mutations on chemotaxis

Mutant	EC_{50}[a]	Efficacy[b]
	nM	%
Wild-type	7.9	100
R8Q/R12Q	960	110
R8Q	490	100
R12Q	35	92
F13A/F14A	21	87
E15Q/H17N	13	110
T31G/A33G/A35G	18	81
H25N/K27Q/R41Q/R47Q/Q48N	4.1	66

[a] EC_{50} was defined as the concentration of CXCL12 that gave 50% of the maximum response as determined by a concentration-response curve for each mutant.
[b] Efficacy was calculated as the ratio of the maximum response of each mutant to the maximum response of wild-type CXCL12 multiplied by 100.

Effect of GAGs on CXCL12-mediated Chemotaxis—As shown in Fig. 2, *A* and *B*, both heparan sulfate and high and low molecular weight heparin caused a clear reduction of the chemotactic activity of CXCL12 in all CXCL12 and GAG concentrations tested. The effect was significant at 10 or 100 μg/ml of GAG for

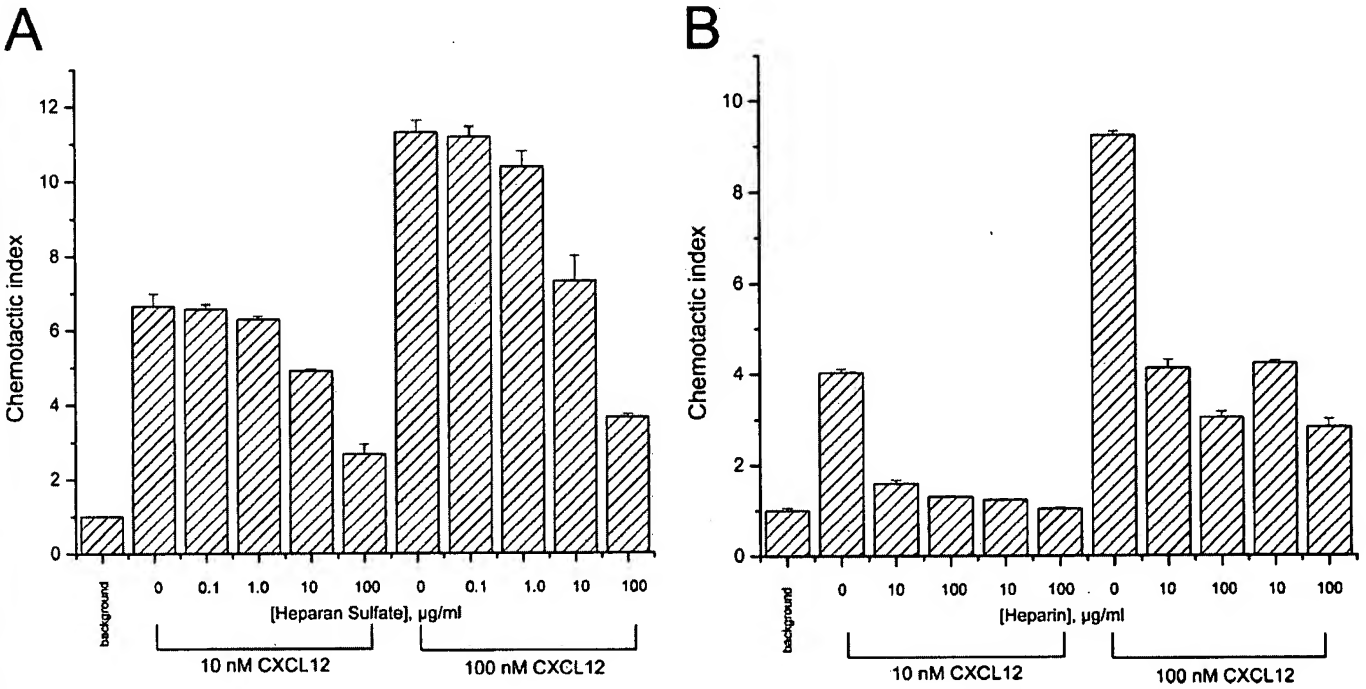


FIGURE 2. **Effects of GAGs on CXCL12-mediated chemotaxis.** Chemotactic index is calculated as the number of cells migrating in the presence of CXCL12 with or without GAGs ÷ number of cells migrating in the absence of CXCL12. *A*, chemotaxis of heparan sulfate (0.1 to 100 μg/ml) preincubated with CXCL12 (10 and 100 nM). *B*, chemotaxis of low and high molecular weight heparin (10 and 100 μg/ml) were preincubated with CXCL12 (10 and 100 nM). Analysis of variance for data gives significance at $p = 0.0001$. In *A*, post hoc *t* tests give $p \leq 0.0008$ for the CXCL12 activity at 10 and 100 nM *versus* background. These activities are inhibited by all concentrations of heparan sulfate tested and the effect is significant for 10 nM ($p \leq 0.009$) and 100 nM CXCL12 ($p \leq 0.009$). In *B*, post hoc *t* tests give $p \leq 0.0003$ for the CXCL12 activity at 10 and 100 nM *versus* background. Both of these activities are significantly inhibited by all concentrations of heparin tested (10 nM CXCL12 is inhibited by 10 and 100 μg/ml of low molecular weight heparin with $p \leq 0.0005$ and by 10 and 100 μg/ml of high molecular weight heparin with $p = 0.0002$ in both cases). 100 nM CXCL12 is inhibited by 10 and 100 μg/ml of low molecular weight heparin with $p \leq 0.0003$ and by 10 and 100 μg/ml of high molecular weight heparin with $p \leq 0.0002$.

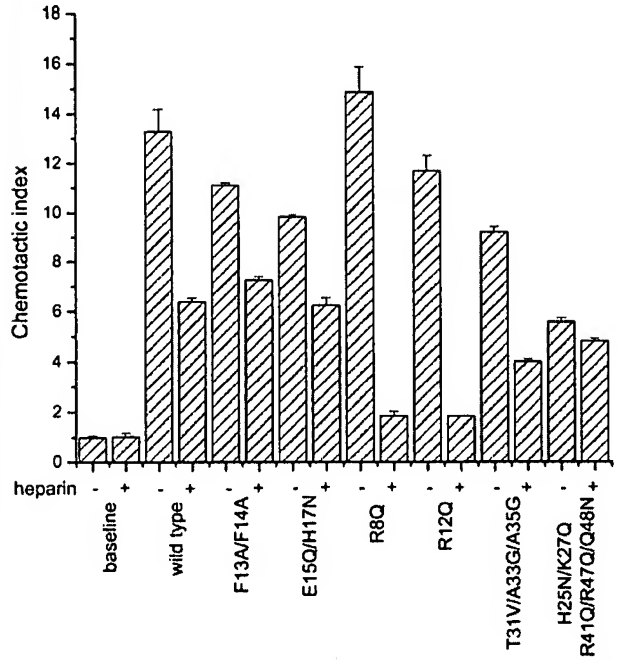


FIGURE 3. **Effects of GAGs on mutant CXCL12-mediated chemotaxis.** Six mutants of CXCL12 were compared with wild-type. Samples with heparin were treated with 100 μg/ml of low molecular weight heparin. All proteins (wild-type and mutants) were tested at an optimal concentration for chemotaxis of CCRF-CEM cells (100 nM) with the exception of R8Q (1.0 μM). The activities of all proteins are significantly higher *versus* background ($p \leq 0.001$). The activities of all proteins (with the exception of the quintuple mutant; $p = 0.02$ for both with and without heparin) are significantly inhibited by treatment with 100 μg/ml of low molecular weight heparin ($p \leq 0.005$).

both 10 and 100 nM CXCL12. These data are in agreement with previous observations on the negative regulatory effects of soluble GAGs on *in vitro* binding of CXCL12 to various cell lines (40, 41) and are consistent with a sequestration mechanism that prevents the interaction of the chemokine with the cell surface. It is important to clarify that these data do not preclude the positive regulatory role of native GAGs on CXCL12 activity and proper presentation to its receptor CXCR4. In fact, treatment of cells with enzymes known to degrade GAGs (heparitinases) induces a significant reduction of CXCL12 binding to cells (40, 41) indicating an important role of native membrane-bound GAGs in binding CXCL12 and increasing its local availability for interaction with CXCR4. It is also noteworthy that *soluble* GAGs have similar inhibitory effects on the activities of most other chemokines tested up to now (21–23) with the exception of the positive effect of heparan sulfate on neutrophil responses to CXCL8 (interleukin-8) (42). It is likely that the soluble heparins that inhibit chemotaxis either do not form a receptor-activating complex with the chemokine or compete with the membrane-bound glycosaminoglycans on the cell that would result in an activated complex.

Effect of Heparin on the Activity of CXCL12 Mutants—Most CXCL12 mutants tested showed significant heparin-induced inhibition of their chemotactic activity (Fig. 3). The F13A/F14A and E15Q/H17N double mutants and the T31V/A33G/A35G triple mutant showed a level of inhibition comparable with that of the wild-type protein indicating that the GAG-chemokine interaction was not significantly modified in these mutants. On the contrary, the H25N/K27Q/R41Q/R47Q/Q48N quintuple mutant showed a clear resistance to inhibition by heparin indicating that the GAG-CXCL12 interaction has been significantly disrupted. Four positive charges are neutralized in this mutant and the data support the suggestion that some or all of these residues, His^{25}, Lys^{27}, Arg^{41}, Arg^{47} (and Gln^{48}), are somehow involved in GAG binding. This is in agreement with previous mutagenesis studies (26). The most unexpected observation from this series of experiments came from the effect of heparin on the R8Q and R12Q single mutants. In both of these cases, the mutants showed a remarkable hypersensitivity to heparin with more than 80% of the activity lost upon binding a GAG *versus* only about 50% for the wild type. These data suggest that the absence of these arginines allows the formation of a GAG-CXCL12 complex with a conformation that increases inhibition.

Effect of Titration of Heparin Disaccharide I-S into a Solution of CXCL12 Monitored by NMR Spectroscopy—To investigate the interactions between CXCL12 and heparin, we used NMR spectroscopy to monitor chemical shifts affected by the addition of an unsaturated heparin disaccharide. ^{1}H-^{15}N-HSQC spectra were collected from a solution of 2.0 mM CXCL12 and increasing concentrations of disaccharide (0–12.0 mM). Fig. 4*A* shows the ^{1}H-^{15}N-HSQC spectrum of 2.0 mM apo-^{15}N-CXCL12 with five overlaid spectra of the CXCL12:heparin disaccharide mixtures. The ratios of CXCL12:heparin disaccharide were 1:0.5, 1:1, 1:2, 1:4, and 1:6. For each resonance peak observed in the apo-CXCL12 spectra, a corresponding peak occurs in each spectrum from the titration, indicating that the overall structure of CXCL12 is maintained. Peaks due to the five arginine side chain δ ^{1}H-^{15}N nuclei were aliased into the spectra at ~121 ppm. Only the side chain of Arg^{41} could be definitively assigned. The spectra from the titration overlay well with the apo spectrum. However, there were a variety of peaks of interest showing a concentration-dependent change in resonance positions. Fig. 4*B* shows a comparison of the absolute change in NMR chemical shift for each residue at a ratio of 2:1 disaccharide:CXCL12 compared with the apo spectrum. Chemical shift perturbations occur in the vicinity of residues identified by mutagenesis to be required for heparin binding. This includes a region around His^{25} and Lys^{27} in the first β-strand and a somewhat smaller region around Arg^{41} in the second β-strand. In the three-dimensional structure of CXCL12, these two regions form a localized cluster along the dimer interface.

Increasing the concentration of disaccharide beyond a 1:2 ratio revealed chemical shift perturbations in two additional regions. One includes the series of His^{17}, Val^{18}, and Ala^{19} and the other includes residues Glu^{63} through Asn^{67} in the COOH-terminal α-helix. These two groups of residues cluster in the three-dimensional structure in a region comparable with the binding site of heparin in CXCL8 as determined by NMR spectroscopy (43). Hence, the NMR spectroscopic study indicates that there are two sites of disaccharide interaction, one corresponding to residues identified by mutation in the dimer interface and a second that corresponds to the site of interaction of heparin with CXCL8 (interleukin-8) (43).

Crystal Structure of the CXCL12-Heparin Disaccharide Complex—Attempts to grow crystals of CXCL12 in complex with various lengths of heparin oligosaccharides were unsuccessful. We therefore used the soaking method. Crystals of native CXCL12

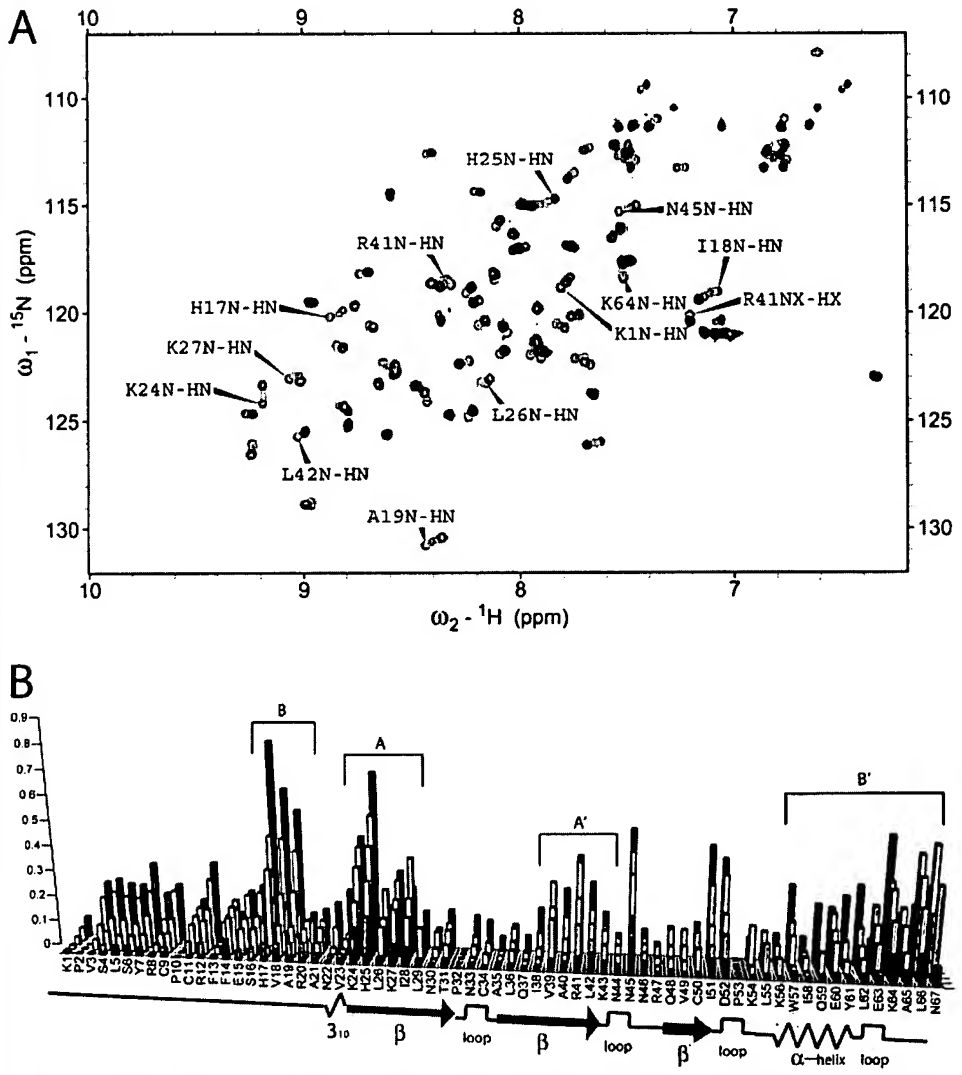


FIGURE 4. **NMR spectroscopic studies of CXCL12 and heparin disaccharide interaction.** *A*, NMR chemical shift changes induced by titration of heparin disaccharide I-S. ^{1}H-^{15}N-HSQC spectra of 2.0 mM ^{15}N-CXCL12 with 1.0, 2.0, 4.0, 8.0, and 12.0 mM heparin disaccharide I-S (*orange*, *yellow*, *light green*, *green*, and *blue*, respectively) are overlaid on the spectrum of 2.0 mM apo ^{15}N-CXCL12 (*red*). Resonance peaks with the largest NMR chemical shift changes are labeled. Peak labels indicate peaks for the apo spectrum. *B*, absolute NMR chemical shift change of each residue for the disaccharide titration. Absolute NMR chemical shift change for each ratio are calculated as $(((|Nppm_{bound} - Nppm_{apo}|) + (|Hppm_{bound} - Hppm_{apo}| \times 10))/2)$. ^{15}N-CXCL12:heparin disaccharide I-S ratios as compared with 1:0 are colored *red* (1:0.5), *orange* (1:1), *yellow* (1:2), *green* (1:4), and *blue* (1:6). Changes in NMR chemical shifts for proline residues are reported as zero as they lack an amide proton. Regions of secondary structure are indicated *below* with *block arrows* representing β regions and *zigzags* representing helical regions. Two sets of residues (*A*/*A*′ and *B*/*B*′) that each form a cluster in the three-dimensional structure that are affected by titration of the disaccharide are indicated by *brackets*.

grew in 4–5 days, measured 400 μm across, diffracted to 2.0 Å, and belonged to the $P2_12_12_1$ space group. Following soaking in a solution containing excess heparin disaccharide, crystals were screened for diffraction and several data sets were collected. Molecular replacement using native CXCL12 (1A15) was successful (4). Composite omit maps were used to view and manually change the positions of the atoms in the protein, where necessary, followed by refinement using REFMAC and the CNS software suite (37, 38). Two disaccharide ligands were modeled into two regions of positive density in a $F_o - F_c$ map.

Analysis of the structure revealed that electron density was not observed for the amino-terminal residues before Ser[4] of chain B. Asn[44] falls into the disallowed region of the Ramachandran plot. Analysis of Asn[44] indicated that the φ and ψ angles are tolerable as they are located in a tight turn in the β-sheet. This residue also has dihedral angles in the disallowed region in the native structures 1A15 (4) and 1QG7 (5). Details of crystallography statistics are available in Table 2.

The crystal structure of the complex is shown in Fig. 5*A*. As in the native structure three β-strands from each subunit form an extended six-stranded β-sheet that in turn forms a pocket where a heparin disaccharide molecule is located. Root mean square deviation values were determined for each α carbon between the apo and bound structures using a local global alignment similarity calculation (44). The average root mean square deviation was 0.85 Å. Regions with significant differences were in chain A residues 1–6 and 29–36 (Fig. 5*B*). (It was not possible to compare the NH_2 terminus of chain B from the CXCL12-disaccharide complex with the native structure because the amino terminus of chain B is not visible in the native structure.)

The disaccharide ligand is visible in the dimer interface nestled within a cluster of basic residues whose side chains are oriented into the pocket (Fig. 6*A*). This ligand is positioned where it had been predicted from mutagenesis studies (26) and is in an orientation that would allow the extension of a polysaccharide chain. CXCL12 directly contacts the sulfate and hydroxyl moieties of the heparin disaccharide. The disaccharide forms hydrogen bonds with His[25] of chain B, Arg[41] of both protein chains, and Lys[27] of chain A (Table 3) (Fig. 6*B*).

A second disaccharide is bound to the exterior of the dimer and hydrogen bonded to Arg[20], Ala[21], and Lys[64] of chain A and Asp[30] of chain B. The location of this second ligand is stabilized by interactions with Val[3] and Ser[4] of chain A from another asymmetric unit, and is consistent with the second site identified by NMR spectroscopy. Table 3 lists the hydrogen bonds observed between the two heparin disaccharides and CXCL12.

DISCUSSION

Heparin and Heparan Sulfate Negatively Affect in Vitro Chemotaxis—The inhibitory effect of soluble GAGs on CXCL12-induced *in vitro* chemotaxis is similar to that previously observed

for other chemokines. This can be explained by a sequestration mechanism that reduces chemokine availability and prevents their binding to their natural interaction partners (membrane-bound receptors and GAGs). The effect of the higher molecular weight GAGs is similar to inhibition of chemokine-mediated inflammation *in vivo* where soluble heparin is injected into the bloodstream (45). We have tested only heparan sulfate and high and low molecular weight heparin, and we assume that the other types of soluble GAGs will induce similar responses. The heparin disaccharide I-S did not exhibit any inhibitory activity (data not shown), presumably due to a low affinity interaction.

TABLE 2
Data collection and refinement statistics

Data collection	
Space group	$P2_12_12_1$
Cell dimensions	
a, b, c (Å)	36.49, 56.97, 71.75
α, β, γ (°)	90, 90, 90
Wavelength (Å)	1.5418
Resolution range[a] (Å)	25.6-2.07 (2.18-2.07)
$I/\sigma I$	7.2 (2.0)
Completeness	99.9% (100%)
R_{merge}[b]	0.057 (0.367)
Redundancy	6.5 (6.5)
Refinement	
Number of reflections	63017 (9045)
Number of unique reflections	9627 (1385)
R_{factor}[c]	0.240
R_{free}[c]	0.265
Number of atoms	1207
Protein	1087
Ligand	70
Water	50
B-factors	
Protein	59.3
Ligand	78.9
Water	55.0
Root mean square deviations	
Bond lengths (Å)	0.025
Bond angles (°)	2.519

[a] Highest resolution shell is shown in parentheses.
[b] $R_{merge} = 100 \times \Sigma_h \Sigma_i |I(h)_i - \langle I(h)\rangle| / \Sigma_h \Sigma_i \langle I(h)_i\rangle$, where I is the observed intensity, and $\langle I\rangle$ is the average intensity of multiple observations of symmetry-related reflections.
[c] $R = \Sigma \| F_o| - |F_c\| / \Sigma |F_o|$. R_{factor} and R_{free} were calculated using the working and test reflection sets, respectively. 5% of the entire reflection was randomly taken as a test set.

NMR Spectroscopy Identified Two Potential Heparin Binding Sites—Positively charged residues clustered in the dimer interface were reported as being required for CXCL12-heparin interaction (26). Our results indicate that residues surrounding His^{25}, Lys^{27}, and Arg^{41} compose one site of interaction in solution as their NMR chemical shifts are perturbed by the disaccharide. A second heparin binding site was observed in solution. This previously unidentified CXCL12 site corresponds to the heparin binding site of CXCL8 (43) and partially corresponds with the heparin binding sites of CXCL4 (platelet factor-4) (46), CXCL1 (growth related protein-α) (46), and CXCL10 (IP-10) (47). Furthermore, there appears to be a difference in the affinities of the two sites for the heparin disaccharide. Chemical shift changes are observed at lower CXCL12: heparin disaccharide I-S ratios for the site in the dimer interface than the region in the amino-terminal loop (residues 17–19).

Two Unsaturated Heparin Disaccharide Binding Sites Are Observed in the Crystal Structure—In the crystal structure one disaccharide is bound at the dimer interface and a second is bound to the amino-terminal loop and the α-helix. Much longer glycosaminoglycan chains are present *in vivo*. This disaccharide contains all the structural elements that are repeated in a longer oligomer with the exception of having an unsaturated hexuronate. With structural data now available we can make further conclusions about the role of amino acids previously thought to interact with heparin. Of the three basic residues in the BB*X*B sequence located in the first β-strand, His^{25} and Lys^{27} were found both to be hydrogen bonded to the disaccharide in the crystal structure. Computer modeling predicted Arg^{41} to bind heparin (28) and we found that Arg^{41} from both chains does indeed interact with the disaccharide. Lys^{43}, on the second β-strand, was predicted by modeling to be important, yet mutagenesis of this single residue had no effect on binding (28). Although Lys^{43} does not interact with the disaccharide in our crystal structure, this does not rule out such interactions with longer glycosaminoglycans *in vivo*. A comparison of the bound unsaturated disaccharide to the previously published molecular model shows that

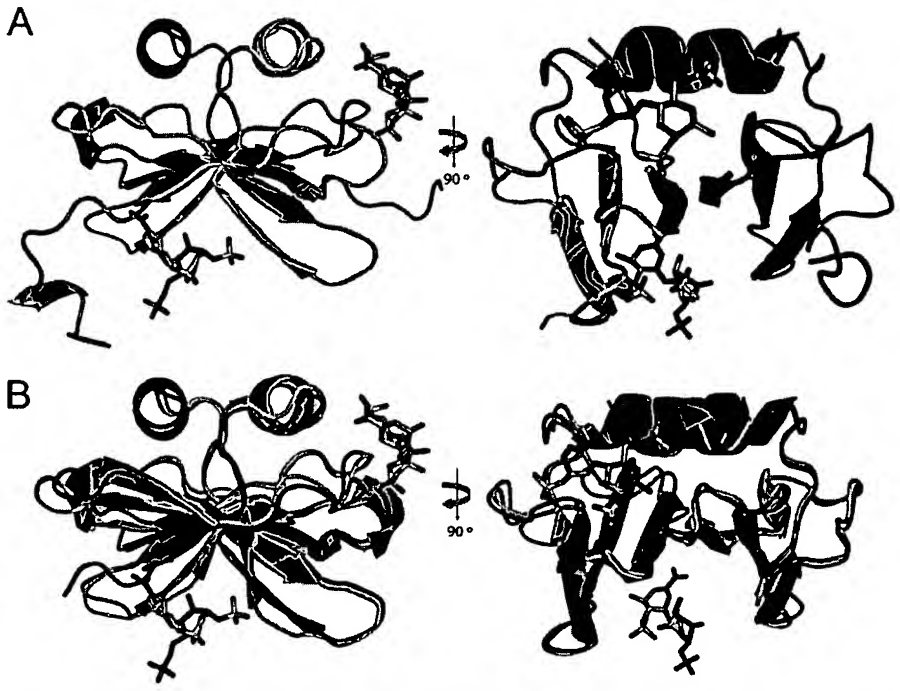


FIGURE 5. **Crystal structure of the CXCL12:heparin disaccharide I-S complex.** *A*, schematic representation of CXCL12 with chain A in *cyan* and chain B in *red*. Two heparin disaccharide I-S ligand molecules are shown as *sticks*, one within the dimer interface and one on the outer portion of monomer A. Image on the *right* is a 90° rotation around the *y* axis. *B*, the disaccharide bound structure of CXCL12 (*blue*) is overlaid with the unbound form (PDB code 1A15) (4) (*cyan*). The amino termini are removed for clarity. Both are shown as schematics representing the secondary structure elements and the two heparin disaccharide I-S molecules are shown as *sticks*. The image on the *right* is a 90° rotation around the *y* axis.

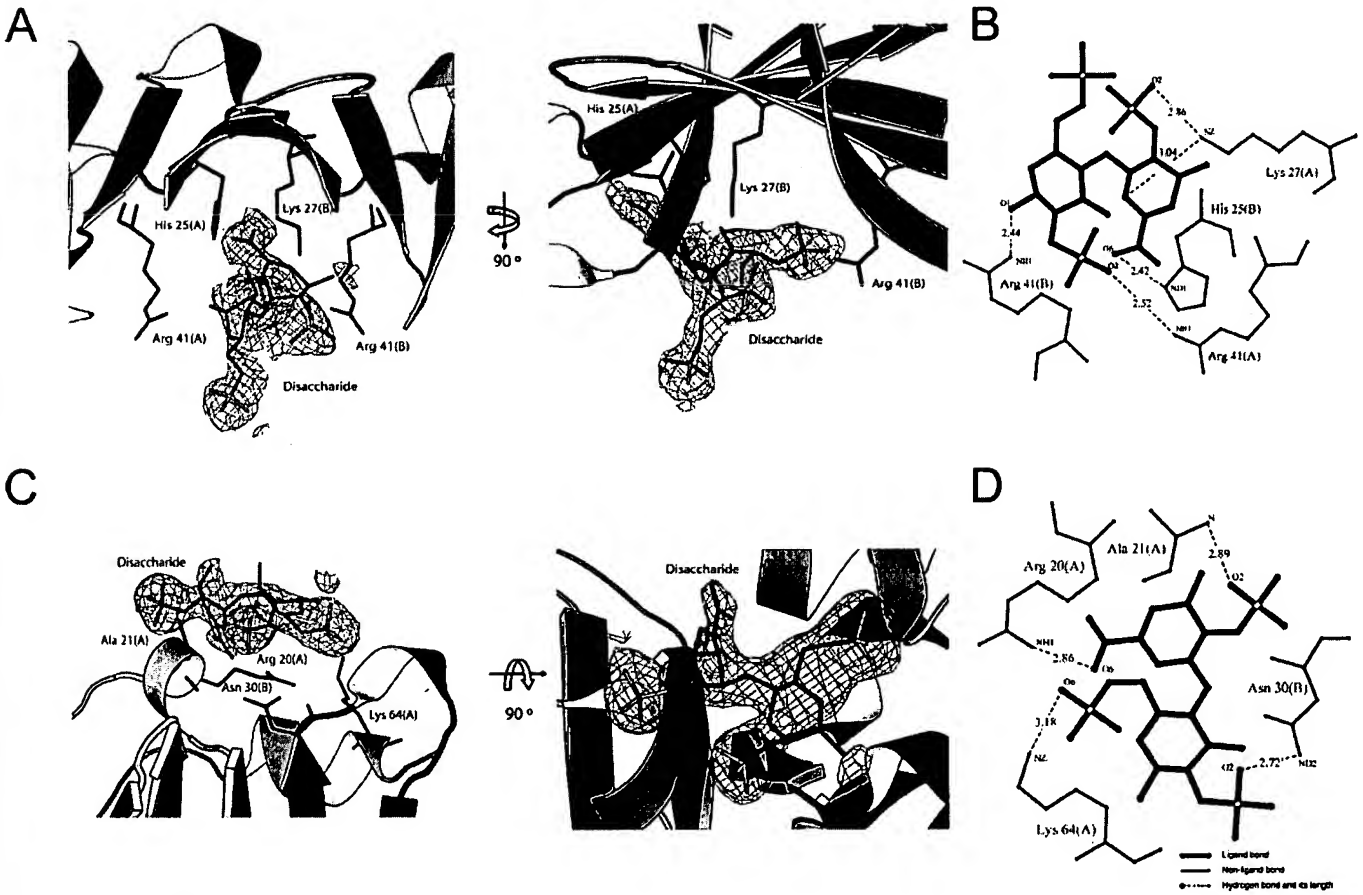


FIGURE 6. **Heparin disaccharides I-S bound to CXCL12.** *A*, the dimer interface from two views rotated 90° of CXCL12 showing the bound disaccharide with a composite omit $2F_o - F_c$ electron density map around the ligand. Side chains interacting with I-S are shown as *sticks*. *B*, the diagram indicates hydrogen bonds (*dashed lines*) between the heparin disaccharide and the CXCL12 residues as calculated by the program LIGPLOT (55). *C*, outer portion of CXCL12 dimer from two views rotated 90° showing the second disaccharide with the corresponding composite omit $2F_o - F_c$ electron density map. *D*, interactions between the second heparin disaccharide and CXCL12.

TABLE 3
Hydrogen bonds between CXCL12 and heparin I-S

Donor, Protein Residue (Chain)	Acceptor			
	Atom	Ligand	Atom	Distance
				Å
His^{25} (B)	ND1	HIS-1[a]	O61(1)	2.42
Lys^{27} (A)	NZ	HIS-1	O-S2(1)	2.86
Lys^{27} (A)	NZ	HIS-1	O5(1)	3.04
Arg^{41} (B)	NH1	HIS-1	O1(2)	2.44
Arg^{41} (A)	NH1	HIS-1	O-S2(2)	2.52
Arg^{20} (A)	NH1	HIS-2	O61(1)	2.86
Ala^{21} (A)	N	HIS-2	O-S2(1)	2.89
Asn^{30} (B)	ND2	HIS-2	O-S2(2)	2.72
Lys^{64} (A)	NZ	HIS-2	O-S6(2)	3.18

[a] HIS-1 and HIS-2 are heparin disaccharides I-S molecules that are bound to the dimer interface and the interleukin-8-like site, respectively.

the disaccharide is oriented slightly differently. This difference in orientation is most likely due to the fact that the disaccharide is small compared with the long glycosaminoglycan chains that were previously modeled (28). Additionally, as the disaccharide is unsaturated, it could bind in a slightly different conformation than a natural ligand.

The second disaccharide binds to the amino-terminal loop and α-helix and forms hydrogen bonds to residues Ala^{20}, Arg^{21}, Asn^{30}, and Lys^{64}. This disaccharide is oriented such that continuation of the polysaccharide is possible without steric conflict. These results indicate that there are multiple sites of interaction on chemokines leading to an avidity effect, resulting in high affinity binding with long chain heparins.

It is interesting to note that there is a difference in the positions of chain A residues 1–6 and 29–36 as compared with the native structure. These two regions are linked by a disulfide bond. Whether heparin binding induces a relay of CXCL12 conformational changes on CXCR4, as suggested by the comparative analysis, to induce signaling by the NH_2 terminus remains to be determined.

Fig. 7 summarizes the structural data from these experiments. The crystal structure of the complex is displayed and regions that are most affected by disaccharide titration monitored by NMR spectroscopy are highlighted. This three-dimensional view gives information about the relative locations of the two disaccharide binding sites in CXCL12.

Comparison of the Heparin Binding Sites of CXCL12 to Those of Other Chemokines—The GAG binding sites of chemokines have been identified by mutagenesis (47), NMR (43, 46), and one crystal structure (48). For CXC chemokines, the heparin binding site of CXCL8 was identified as a cluster of residues with changes in chemical shift upon addition of heparin disaccharide I-S (43). This site is similar to the second site of CXCL12 and involves both CXCL8 residues 18–23 in the loop

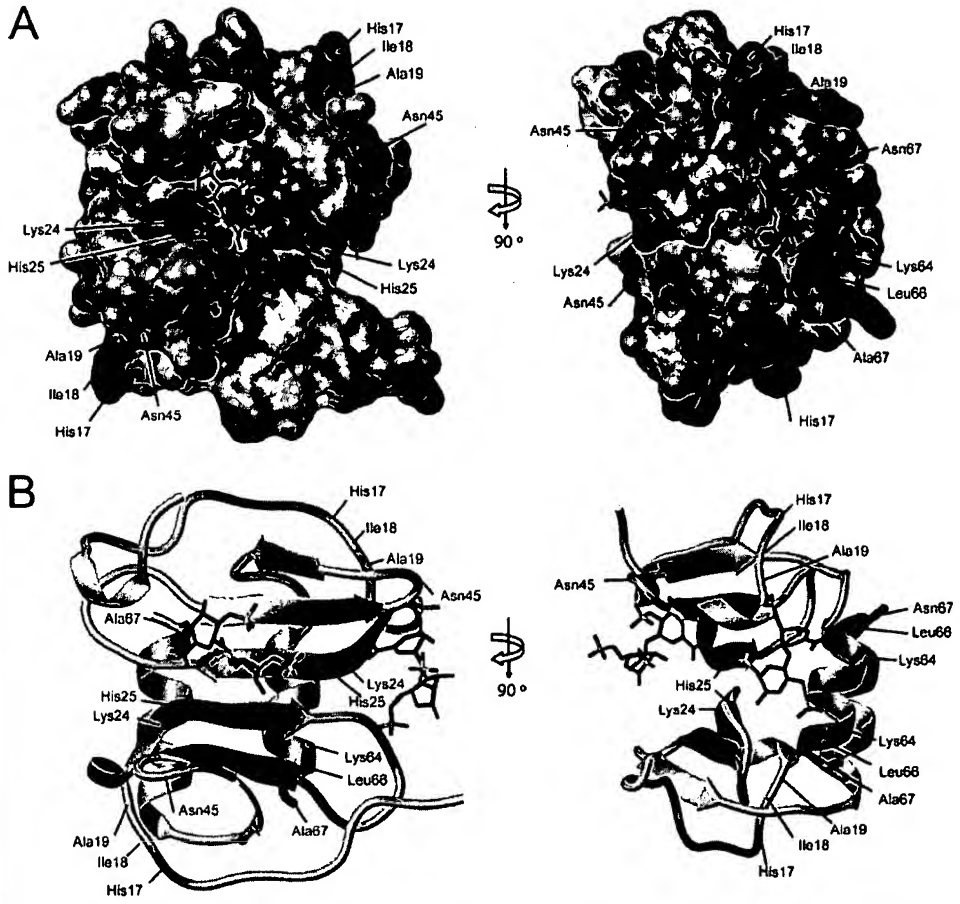


FIGURE 7. **An overlay of the results from the backbone chemical shifts changes upon heparin disaccharide addition on the crystal structure of the CXCL12-heparin disaccharide I-S complex.** *A*, calculated Connolly surface of CXCL12 with heparin disaccharide ligands displayed as *sticks*. Greater chemical shift changes are displayed as increasing *red* intensity. Chemical shift changes are measured between apo and a 1:2 CXCL12:heparin disaccharide I-S molar ratio. *Light gray* indicates prolines for which chemical shift data are not available. *B*, the carbon backbones are displayed as schematics indicating the secondary structure and colored as in *A*. The figures on the *right* are rotated 90° around the *y* axis.

preceding the first β-strand and residues from the COOH-terminal α-helix. The interactions between a heparin dodecasaccharide and CXCL4 and CXCL1 were also determined using NMR spectroscopy (46). Like other chemokines, basic residues are involved and, as in CXCL8, chemical shift changes for residues at the loop immediately before the first β-strand implicate this region in contacts with the heparin oligomer. The heparin binding site of CXCL10 was determined using mutagenesis, binding experiments, and *in vitro* cellular assays (47). Our conclusion from this analysis is that CXCL12 has a unique heparin binding site in the dimer interface involving β-strands one and two in addition to a more common heparin binding site involving the loop region preceding the first β-strand and residues from the α-helix. Our interpretation is that there may be sites of varying affinities for disaccharides, each of which provides an avidity effect leading to high affinity interactions with a long chain GAG that binds CXCL12, inducing dimerization, and sequestering CXCL12.

The heparin binding sites of numerous CC chemokines have also been identified (27, 48–50). CC chemokines, with few exceptions, dimerize with a different topology and it would be interesting to compare the GAG binding sites. CCL5 (RANTES (regulated on activation normal T cell expressed and secreted)) was co-crystallized with heparin disaccharides (I-S and III-S) (48). In the CCL5-disaccharide complexes, a single disaccharide is bound at different locations to three crystallographically symmetric dimers of CCL5. Longer chain heparin molecules may have interactions with residues at all these locations. None of these three sites of interaction are similar to the CXCL12 sites of interaction we have identified, with the exception of His^{23} in CCL5 (corresponding to His^{25} in CXCL12). The CC and CXC chemokines appear to be different not only in their dimeric structures, but also in the location of the primary, tertiary, and quaternary sites involved with interactions with GAGs. The vast varieties of substituted glycosaminoglycans present *in vivo* may be responsible for selectivity of receptor binding for chemokines in different families and also within families.

One Binding Site of Heparin Partially Overlaps the Binding Site of the NH_2-terminal Peptide of CXCR4—The two-stage model of chemokine-chemokine receptor activation dictates an initial binding event followed by a conformational change leading to activation (51–53). Clore and coworkers (33) completed a study in which a peptide equivalent to residues 1–27 of CXCR4 was added to CXCL12 in solution. It was found that the pocket formed by the β-strands in the dimer interface, which includes residues Lys^{24}, His^{25}, Ala^{40}, Arg^{41}, Gln^{48}, and Tyr^{61} of CXCL12, interacts with the CXCR4 peptide (33). This region is part of the heparin binding site. This supports a proposed mechanism for chemokine presentation by proteoglycans whereby the chemokine GPCR displaces proteoglycans bound to the chemokine agonist (22). Initially, secreted CXCL12 would be bound to GAGs attached to proteoglycans via multiple sites on CXCL12. This binding would increase dimerization of CXCL12 (8), protect CXCL12 from inactivation by CD26 (24), and increase the localized concentration of CXCL12. GAGs would then be displaced from CXCL12 by the amino-terminal region of CXCR4. CXCR4 could then be activated by the amino terminus of CXCL12 (6). Interestingly, the binding of GAGs to CXCL10 (47) and CCL2 (49) are also partially overlapping with their respective receptor binding sites. This suggests the chemokines from different families use similar mechanisms with different heparin binding sites.

Therapeutic intervention of diseases involving chemokines and their GPCRs traditionally targets the receptors with antagonists or otherwise modulates the signaling at the level of the

receptor. Due to complications arising from multiple chemokines activating a single receptor and vice versa, targeting the initial glycosaminoglycan-chemokine interaction with small GAG-derived molecules may prove to be an alternative avenue for disease therapy. For example, therapeutic pentasaccharide heparin is often used clinically, functioning as an anticoagulant as well as an anti-inflammatory agent (54). Likewise, small heparin-based molecules may function as a potential new class of anti-inflammatory compounds that do not stimulate anticoagulant activity. Our structure of the disaccharide-bound dimer can serve as the starting point for the design of such small molecule inhibitors of CXCL12-GAG interaction.

Acknowledgments—We thank Gregg Crichlow and Paul Pepin for expert assistance with the crystallography, Sharon Mella for assistance with tissue culture, Demetrios Braddock for assistance with protein purification, and Camille Keeler for assistance with the NMR spectroscopy. We thank Dr. Lortat-Jacob for providing atomic coordinates of the modeled CXCL12-heparin tetradecasaccharide complex.

REFERENCES

1. Fernandez, E. J., and Lolis, E. (2002) *Annu. Rev. Pharmacol. Toxicol.* **42,** 469–499
2. Lapidot, T., and Petit, I. (2002) *Exp. Hematol.* **30,** 973–981
3. Bleul, C. C., Farzan, M., Choe, H., Parolin, C., Clark-Lewis, I., Sodroski, J., and Springer, T. A. (1996) *Nature* **382,** 829–833
4. Dealwis, C., Fernandez, E. J., Thompson, D. A., Simon, R. J., Siani, M. A., and Lolis, E. (1998) *Proc. Natl. Acad. Sci. U. S. A.* **95,** 6941–6946
5. Ohnishi, Y., Senda, T., Nandhagopal, N., Sugimoto, K., Shioda, T., Nagal, Y., and Mitsui, Y. (2000) *J. Interferon Cytokine Res.* **20,** 691–700
6. Crump, M. P., Gong, J. H., Loetscher, P., Rajarathnam, K., Amara, A., Arenzana-Seisdedos, F., Virelizier, J. L., Baggiolini, M., Sykes, B. D., and Clark-Lewis, I. (1997) *EMBO J.* **16,** 6996–7007
7. Holmes, W. D., Consler, T. G., Dallas, W. S., Rocque, W. J., and Willard, D. H. (2001) *Protein Expr. Purif.* **21,** 367–377
8. Veldkamp, C. T., Peterson, F. C., Pelzek, A. J., and Volkman, B. F. (2005) *Protein Sci.* **14,** 1071–1081
9. White, F. A., Bhangoo, S. K., and Miller, R. J. (2005) *Nat. Rev. Drug Discov.* **4,** 834–844
10. Kucia, M., Jankowski, K., Reca, R., Wysoczynski, M., Bandura, L., Allendorf, D. J., Zhang, J., Ratajczak, J., and Ratajczak, M. Z. (2004) *J. Mol. Histol.* **35,** 233–245
11. Nagasawa, T., Tachibana, K., and Kishimoto, T. (1998) *Semin. Immunol.* **10,** 179–185
12. Kucia, M., Reca, R., Miekus, K., Wanzeck, J., Wojakowski, W., Janowska-Wieczorek, A., Ratajczak, J., and Ratajczak, M. Z. (2005) *Stem Cells (Durham)* **23,** 879–894
13. Kucia, M., Ratajczak, J., and Ratajczak, M. Z. (2005) *Biol. Cell* **97,** 133–146
14. Bajetto, A., Bonavia, R., Barbero, S., Florio, T., and Schettini, G. (2001) *Front. Neuroendocrinol.* **22,** 147–184
15. Zhou, N., Luo, Z., Luo, J., Liu, D., Hall, J. W., Pomerantz, R. J., and Huang, Z. (2001) *J. Biol. Chem.* **276,** 42826–42833
16. Kaul, M., and Lipton, S. A. (2006) *Curr. HIV Res.* **4,** 307–318
17. Muller, A., Homey, B., Soto, H., Ge, N., Catron, D., Buchanan, M. E., McClanahan, T., Murphy, E., Yuan, W., Wagner, S. N., Barrera, J. L., Mohar, A., Verastegui, E., and Zlotnik, A. (2001) *Nature* **410,** 50–56
18. Luker, K. E., and Luker, G. D. (2006) *Cancer Lett.* **238,** 30–41
19. Epstein, R. J. (2004) *Nat. Rev. Cancer* **4,** 901–909
20. Diaz, G. A., and Gulino, A. V. (2005) *Curr. Allergy Asthma Rep.* **5,** 350–355
21. Handel, T. M., Johnson, Z., Crown, S. E., Lau, E. K., and Proudfoot, A. E. (2005) *Annu. Rev. Biochem.* **74,** 385–410
22. Lau, E. K., Allen, S., Hsu, A. R., and Handel, T. M. (2004) *Adv. Protein Chem.* **68,** 351–391
23. Johnson, Z., Proudfoot, A. E., and Handel, T. M. (2005) *Cytokine Growth Factor Rev.* **16,** 625–636
24. Sadir, R., Imberty, A., Baleux, F., and Lortat-Jacob, H. (2004) *J. Biol. Chem.* **279,** 43854–43860
25. Bleul, C. C., Fuhlbrigge, R. C., Casasnovas, J. M., Aiuti, A., and Springer, T. A. (1996) *J. Exp. Med.* **184,** 1101–1109
26. Amara, A., Lorthioir, O., Valenzuela, A., Magerus, A., Thelen, M., Montes, M., Virelizier, J. L., Delepierre, M., Baleux, F., Lortat-Jacob, H., and Arenzana-Seisdedos, F. (1999) *J. Biol. Chem.* **274,** 23916–23925
27. Proudfoot, A. E., Handel, T. M., Johnson, Z., Lau, E. K., LiWang, P., Clark-Lewis, I., Borlat, F., Wells, T. N., and Kosco-Vilbois, M. H. (2003) *Proc. Natl. Acad. Sci. U. S. A.* **100,** 1885–1890
28. Sadir, R., Baleux, F., Grosdidier, A., Imberty, A., and Lortat-Jacob, H. (2001) *J. Biol. Chem.* **276,** 8288–8296
29. Foley, G. E., Lazarus, H., Farber, S., Uzman, B. G., Boone, B. A., and McCarthy, R. E. (1965) *Cancer* **18,** 522–529
30. Sachpatzidis, A., Benton, B. K., Manfredi, J. P., Wang, H., Hamilton, A., Dohlman, H. G., and Lolis, E. (2003) *J. Biol. Chem.* **278,** 896–907
31. Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J., and Bax, A. (1995) *J. Biomol. NMR* **6,** 277–293
32. Wallace, A. C., Laskowski, R. A., and Thornton, J. M. (1995) *Protein Eng.* **8,** 127–134
33. Gozansky, E. K., Louis, J. M., Caffrey, M., and Clore, G. M. (2005) *J. Mol. Biol.* **345,** 651–658
34. Leslie, A. G. (1999) *Acta Crystallogr. D Biol. Crystallogr.* **55,** 1696–1702
35. Collaborative Computational Project Number Four (1994) *Acta Crystallogr. D Biol. Crystallogr.* **50,** 760–763
36. McCoy, A. J., Grosse-Kunstleve, R. W., Storoni, L. C., and Read, R. J. (2005) *Acta Crystallogr. D Biol. Crystallogr.* **61,** 458–464
37. Potterton, E., Briggs, P., Turkenburg, M., and Dodson, E. (2003) *Acta Crystallogr. D Biol. Crystallogr.* **59,** 1131–1137
38. Brunger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T., and Warren, G. L. (1998) *Acta Crystallogr. D Biol. Crystallogr.* **54,** 905–921
39. Schuttelkopf, A. W., and van Aalten, D. M. (2004) *Acta Crystallogr. D Biol. Crystallogr.* **60,** 1355–1363
40. Mbemba, E., Benjouad, A., Saffar, L., and Gattegno, L. (1999) *Virology* **265,** 354–364
41. Mbemba, E., Gluckman, J. C., and Gattegno, L. (2000) *Glycobiology* **10,** 21–29
42. Webb, L. M., Ehrengruber, M. U., Clark-Lewis, I., Baggiolini, M., and Rot, A. (1993) *Proc. Natl. Acad. Sci. U. S. A.* **90,** 7158–7162
43. Kuschert, G. S., Hoogewerf, A. J., Proudfoot, A. E., Chung, C. W., Cooke, R. M., Hubbard, R. E., Wells, T. N., and Sanderson, P. N. (1998) *Biochemistry* **37,** 11193–11201
44. Zemla, A. (2003) *Nucleic Acids Res.* **31,** 3370–3374
45. Johnson, Z., Kosco-Vilbois, M. H., Herren, S., Cirillo, R., Muzio, V., Zaratin, P., Carbonatto, M., Mack, M., Smailbegovic, A., Rose, M., Lever, R., Page, C., Wells, T. N., and Proudfoot, A. E. (2004) *J. Immunol.* **173,** 5776–5785
46. Mikhailov, D., Young, H. C., Linhardt, R. J., and Mayo, K. H. (1999) *J. Biol. Chem.* **274,** 25317–25329
47. Campanella, G. S., Lee, E. M., Sun, J., and Luster, A. D. (2003) *J. Biol. Chem.* **278,** 17066–17074
48. Shaw, J. P., Johnson, Z., Borlat, F., Zwahlen, C., Kungl, A., Roulin, K., Harrenga, A., Wells, T. N., and Proudfoot, A. E. (2004) *Structure* **12,** 2081–2093
49. Lau, E. K., Paavola, C. D., Johnson, Z., Gaudry, J. P., Geretti, E., Borlat, F., Kungl, A. J., Proudfoot, A. E., and Handel, T. M. (2004) *J. Biol. Chem.* **279,** 22294–22305
50. Stringer, S. E., Nelson, M. S., and Gupta, P. (2003) *Blood* **101,** 2243–2245
51. Xanthou, G., Williams, T. J., and Pease, J. E. (2003) *Eur. J. Immunol.* **33,** 2927–2936
52. Cai, S. H., Tan, Y., Ren, X. D., Li, X. H., Cai, S. X., and Du, J. (2004) *Acta Pharmacol. Sin.* **25,** 152–160
53. Wells, T. N., Power, C. A., Lusti-Narasimhan, M., Hoogewerf, A. J., Cooke, R. M., Chung, C. W., Peitsch, M. C., and Proudfoot, A. E. (1996) *J. Leukocyte Biol.* **59,** 53–60
54. Tyrell, D. J., Kilfeather, S., and Page, C. P. (1995) *Trends Pharmacol. Sci.* **16,** 198–204

On the Accuracy of Homology Modeling and Sequence Alignment Methods Applied to Membrane Proteins

Lucy R. Forrest, Christopher L. Tang, and Barry Honig
Howard Hughes Medical Institute, Department of Biochemistry and Molecular Biophysics, Center for Computational Biology and Bioinformatics, Columbia University, New York, New York 10032

ABSTRACT In this study, we investigate the extent to which techniques for homology modeling that were developed for water-soluble proteins are appropriate for membrane proteins as well. To this end we present an assessment of current strategies for homology modeling of membrane proteins and introduce a benchmark data set of homologous membrane protein structures, called HOMEP. First, we use HOMEP to reveal the relationship between sequence identity and structural similarity in membrane proteins. This analysis indicates that homology modeling is at least as applicable to membrane proteins as it is to water-soluble proteins and that acceptable models (with Cα-RMSD values to the native of 2 Å or less in the transmembrane regions) may be obtained for template sequence identities of 30% or higher if an accurate alignment of the sequences is used. Second, we show that secondary-structure prediction algorithms that were developed for water-soluble proteins perform approximately as well for membrane proteins. Third, we provide a comparison of a set of commonly used sequence alignment algorithms as applied to membrane proteins. We find that high-accuracy alignments of membrane protein sequences can be obtained using state-of-the-art profile-to-profile methods that were developed for water-soluble proteins. Improvements are observed when weights derived from the secondary structure of the query and the template are used in the scoring of the alignment, a result which relies on the accuracy of the secondary-structure prediction of the query sequence. The most accurate alignments were obtained using template profiles constructed with the aid of structural alignments. In contrast, a simple sequence-to-sequence alignment algorithm, using a membrane protein-specific substitution matrix, shows no improvement in alignment accuracy. We suggest that profile-to-profile alignment methods should be adopted to maximize the accuracy of homology models of membrane proteins.

INTRODUCTION

Membrane proteins are believed to comprise 20–30% of the proteins in a genome (1–3) and represent a significant proportion of therapeutic drug targets (4). However, as a result of difficulties in experimental structure determination, they constitute only ~1% of the structures available in the protein data bank (PDB) (5). The absence of structural information severely limits our ability to understand membrane protein function. Based on previous experience with water-soluble proteins, it is likely that computational structure prediction will provide a useful approach to generating models for these proteins. Typically, the most accurate models of protein structures are achieved through homology modeling, where a known structure is used as a template for the construction of a model of a related protein (6). However, it remains unclear whether the methods and assumptions used in homology modeling of water-soluble proteins can be applied directly to membrane proteins without modification.

There are several features of membrane proteins that distinguish them from water-soluble proteins. The differences arise because the environment of the transmembrane regions of membrane proteins is different from that in aqueous solution: it is predominantly lipophilic, lacks hydrogen-bonding potential, and provides little screening of electrostatic interactions. At the primary sequence level, this results in significant differences in amino acid composition (7,8) and in the probabilities of amino acid substitutions during evolution (9,10), generally favoring residues with hydrophobic side chains, especially at the protein-lipid interface (11,12). In addition, amino acids have been shown to have different secondary-structure propensities in membrane environments and in aqueous solution (13–15).

The differences in the properties of the two types of protein might be expected to have consequences for the applicability of some homology modeling methods to membrane proteins. For example, differences in amino acid composition and evolutionary substitution probabilities imply that methods for the alignment of protein sequences may not be directly transferable. This possibility has led to the creation of novel amino acid substitution matrices (10,16), which are used to identify probable matches in sequences, and to the introduction of so-called bipartite alignment methods that utilize these matrices in transmembrane regions only (10,16,17).

A second aspect of modeling that may be affected by the differences between membrane proteins and water-soluble proteins is the prediction of secondary structure. We draw a distinction between the secondary structure of a residue and its location relative to the membrane, since every amino acid can be labeled as having both a specific secondary-structure type and a specific location. This distinction is useful because it allows for the unique description of secondary-structure elements peripheral to the membrane (18), as well as coil-like residues within the membrane, e.g., in reentrant loops or unwound helices (19). Thus, a method capable of accurately predicting the secondary structure of each residue

Submitted February 28, 2006, and accepted for publication April 13, 2006.
Address reprint requests to Barry Honig, E-mail: bh6@columbia.edu.

0006-3495/06/07/508/10 $2.00
doi: 10.1529/biophysj.106.082313

in a membrane protein sequence would provide information that is supplementary to that obtained from the prediction of the location of a particular amino acid with respect to the bilayer. More generally, it is important to understand the extent to which secondary-structure prediction algorithms designed for soluble proteins are applicable to membrane proteins.

A third way the membrane environment may affect homology modeling studies involves the presence of unique topological constraints provided by the lipid bilayer (20). In principle, it is possible that the range of relative orientations of helices within the membrane is more restricted than in the aqueous phase, which may limit the structural diversity available to families of membrane proteins. It might also suggest that homology models of membrane proteins are more accurate than models of water-soluble proteins for the same level of sequence identity. It is therefore of interest to assess the relationship between sequence identity and structural similarity for membrane proteins.

In this work, we address the three issues raised above. We analyze the performance of state-of-the-art globular-protein homology modeling strategies using a set of 36 homologous membrane protein structures (HOMEP), comprising 11 families of topologically related proteins. Taking each protein in turn, we use all its family members as templates for the construction of homology models whose accuracy is then determined by comparison to the known structure. Although small on the scale of general sequence alignment benchmark sets such as BaliBase (21), the HOMEP set is carefully compiled and covers a wide range of sequence identities, varying from 80 to <10%.

METHODS

The HOMEP benchmark set

A data set of 36 HOMEP structures (see Supplementary Material Table 1; the data set is available at http://trantor.bioc.columbia.edu/~lucy/homep) was selected from the PDB (5). All the proteins were solved using x-ray crystallography at a resolution of 3.5 Å or better. If two or more structures of the same protein were available, that with the highest resolution was selected. Polypeptide chains believed not to contact the membrane were omitted. Each family contains proteins with the same topology, defined here as the number and orientation of the transmembrane domains, excluding peripheral membrane-spanning domains that are not present in all members of the family. Taking each protein as a potential query sequence and all other members of its family as templates (for a homology model), the HOMEP data set contains 94 query-template pairs, from which 94 alignments and homology models can be constructed (Supplementary Material Table 2).

Two definitions of the transmembrane regions were adopted. The first, referred to as TM, was defined by hand to incorporate all residues in membrane-spanning secondary-structure elements according to DSSP (22) that were also superimposed in the structural alignment of all family members. Thus, the TM regions include residues located at the lipid-water interface as well as within the bilayer (Supplementary Material Table 3). The second definition, referred to as TMDET, comprises only residues in the hydrophobic core of the membrane, as defined by the TMDET algorithm (23) used by the PDB_TM database (24). Two short segments were incorrectly assigned by TMDET and thus excluded from the analysis: a strand (residues 128–133) in a loop region of 1osm and a helical region in the first two N-terminal residues of 1pw4.

Secondary-structure prediction accuracy

Since HOMEP is highly redundant by design, for the analysis of secondary-structure prediction algorithms we used the 40% nonredundant set of membrane proteins from the PDB_TM database from July 1, 2005. After excluding theoretical models, Cα-only structures, and proteins with missing residues, the set contained 106 chains from 71 membrane proteins, of which 92 chains were α-helical and 14 chains were β-barrels. Predictions were obtained with local installations of PSIPRED (25) v2.3, JNET (26), and PHDsec (27), and compared against assignments from DSSP. To obtain the multiple-sequence alignment input for each protein, we ran a PSI-BLAST search on the National Center for Biotechnology Information (NCBI) nonredundant database (*nr*); we ran three PSI-BLAST iterations including sequences below an E-value cutoff of 5×10^{-4} and reported sequences with an E-value cutoff of 1×10^{-3}. No filtering of transmembrane regions was carried out.

We also assessed the composite prediction used by HMAP (28), which is a vector of probabilities for the three states (helix, strand, and coil) determined by direct averaging of the confidence scores from PSIPRED, JNET, and PHDsec. To enable comparison with the DSSP assignments, the prediction at each position was taken as the state with the highest probability.

Generation of sequence alignments

Sequence-to-sequence alignments

The dynamic programming algorithm in ClustalW v1.82 (29) was used to align each of the query-template sequence pairs. Gap-open penalties (p_o) of 9, 10, 11, 12, 15, and 20 were tested in combination with gap-extension penalties (p_e) of 0.1 or 1. No clear difference was seen in the Q or ALO scores (see below) of pairwise alignments using these different gap penalties (data not shown), so the default values ($p_o = 10$ and $p_e = 0.1$) were used.

Sequence-to-profile alignments

We carried out PSI-BLAST (30) searches for each template sequence on the *nr* database, which was clustered at 65% sequence identity; five iterations of PSI-BLAST were carried out using E-value cutoffs as above. The sequence hits were compiled into a multiple-sequence alignment from which very remote homologs were removed according to the sequence threshold of Batalov and Abagyan as described by Tang et al. (28). This purged alignment was then used to create a sequence-based profile to which the query sequence was aligned with ClustalW, creating a sequence-to-profile alignment. A profile is an alternate representation of the primary sequence in which each amino acid position contains a set of probabilities.

Multiple-sequence alignments

These were generated by combining PSI-BLAST hits (as above) for both query and template into a single nonredundant set of sequences, which were then aligned using ClustalW, (T-Coffee (31), Muscle (32), and ProbCons (33)).

HMAP profile-to-profile alignments

HMAP is a program for the construction and alignment of structure-based profiles (28) that is similar in its algorithms to other profile-based approaches (34). For each template we generated two types of profile: HMAP [1,2] and HMAP [1,2,3], which combine sequence and secondary- and tertiary-structure information in different ways. The HMAP [1,2] template profiles combined sequence information from a PSI-BLAST search (as above) with a consensus secondary-structure assignment derived from all templates in the family, alongside position-specific weights reflecting the location of ungapped (i.e., core) positions in the alignment. The HMAP [1,2,3] template profiles differ in that the PSI-BLAST hits were taken from all available templates and merged using a structural alignment as a guide. For the query

sequence we created a similar HMAP [1,2] profile, except that the secondary structure was obtained from a consensus prediction (see above) and the position-specific weights depended on the confidence levels of those predictions. Query and template profiles were then aligned using a score designed to favor matching of ungapped core regions and of secondary-structure types. Gap penalties were also assigned according to the location of core regions or secondary-structure elements. We used the local-global alignment method where unaligned terminal residues are only penalized in the query.

In the case of the reductase family of proteins, one member (PDB code: 1l0v) comprises two protein chains, whereas the homologous region in the other two reductase proteins is made up by a single chain. Alignment therefore required concatenation of the sequences or profiles of the two 1l0v chains; multiple sequence alignments were not possible.

Structure-based alignments

Structure-based sequence alignments were carried out with SKA (35,36). Residues that were matched in the structure alignment were used to define the correct alignment, which is the reference state in the calculation of the percentage of aligned positions that are correctly predicted, Q (see below). The sequence identity for each query-template pair was calculated using this alignment and was defined as the number of identical residues divided by the length of the shortest sequence.

Measures of accuracy

Models were built using Modeller 6v2 (37) and were assessed using several measures of structure similarity or model accuracy. In addition to the root mean squared deviation of the positions of the Cα atoms (Cα-RMSD), we compare the model with the native structure using two scores that are used to evaluate predictions in CASP (38). Both measures are based on the global distance test (GDT), which determines the number of model-template Cα-atom pairs, G(v) that are within a distance threshold, v Å (39). Using GDT results, the GDT_TS score (40) is then calculated as the average percentage of residues that fit within four different cutoff distances:

$$\mathrm{GDT\_TS}(\%) = \frac{1}{4}\sum_{v=1,2,4,8}\left[\frac{G(v)}{t}\times 100\right],$$

where t is the number of Cα-atoms in the template structure. A second measure, the AL0 score (37), is computed in a similar way but using a single threshold of 3.8 Å, that is

$$\mathrm{AL0}(\%) = \frac{G(3.8)}{t}\times 100.$$

This threshold corresponds approximately to the distance between adjacent Cα atoms in a peptide chain, so that it tends to reflect structural differences corresponding to shifts in the sequence alignment.

Sequence alignment accuracy was also measured using the percentage of correctly aligned positions, Q:

$$\mathrm{Q}(\%) = \frac{N_c}{N_a}\times 100,$$

where N_a is the number of nongapped positions in the structure-based SKA alignment and N_c is the number of correctly aligned positions in the test alignment compared to the SKA alignment.

For ease of comparison, the individual membrane protein models in our set (one for each query-template pair, M, have been ranked according to i) the fraction of the target structure that can be superimposed on the template within a cutoff distance of 5 Å, and ii) the sequence identity between the target and template. These two rankings, respectively denoted by R^f_M and R^i_M, were combined into a relative difficulty score (41) for each model: Difficulty(M) = $(R^f_M + R^i_M)/2$.

RESULTS

Benchmark of membrane protein homology model accuracy

For each of the 94 pairs of membrane proteins in the HOMEP data set, a homology model was built using the structure-based sequence alignment, which we take as the correct alignment. The Cα-RMSD and GDT_TS scores of these models, plotted against sequence identity (Fig. 1), provide a benchmark of the likely quality of a membrane protein homology model for a given level of sequence identity, assuming that the correct alignment can be achieved and that no refinement is carried out. Fig. 1 shows that the quality of a membrane protein homology model decreases exponentially with decreasing sequence identity.

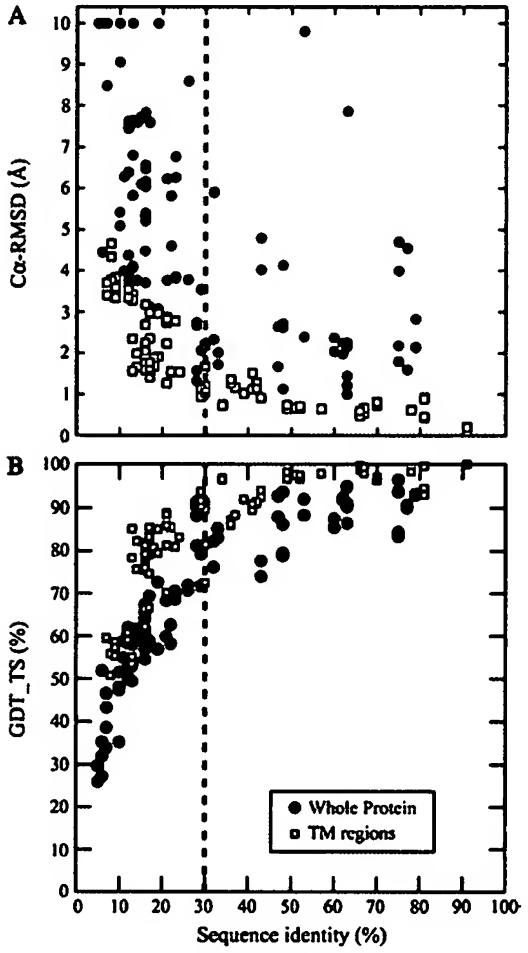


FIGURE 1 Structural relationship between membrane protein models and their templates. The sequence identity of the structure-based (correct) alignment is plotted against (A) the Cα-RMSD and (B) the GDT_TS scores of the corresponding model compared to the native structure. Data are shown for the whole protein (●) and for the transmembrane regions (□). Six models had RMSD values of between 10 and 40 Å; for clarity these points are plotted at RMSD = 10 Å.

Since the alignments used to generate these homology models are based on structural (i.e., optimal) alignments, Fig. 1 also contains information on the structural similarity between the target and query crystal structures. As such, the exponential relationship between sequence and structure for these membrane proteins appears to be very similar to that observed for pairs of homologous water-soluble proteins (42–44). The TM definition used here corresponds loosely to the common core defined by Chothia and Lesk (44); the Cα-RMSD values of the two data sets match reasonably well. The membrane protein whole-protein Cα-RMSDs are more similar to the values of Flores et al. (43), which also represent whole proteins, although this comparison is more difficult due to the large number of outliers in our data set. These outliers are caused by the absence of template regions for certain long (>10 residue) loops and termini, resulting in large local errors to which the RMSD measure is particularly sensitive. When AL0 and GDT_TS scores are used, however, it is clear that the scores for the whole models are indeed significantly lower than the scores for the transmembrane regions (Fig. 2). This suggests that there is a marked structural variability in the connecting regions between membrane-spanning segments of topologically related proteins (i.e., with the same number of transmembrane domains and the same N- to C-terminal orientations), as indicated by the variability in their length and sequences.

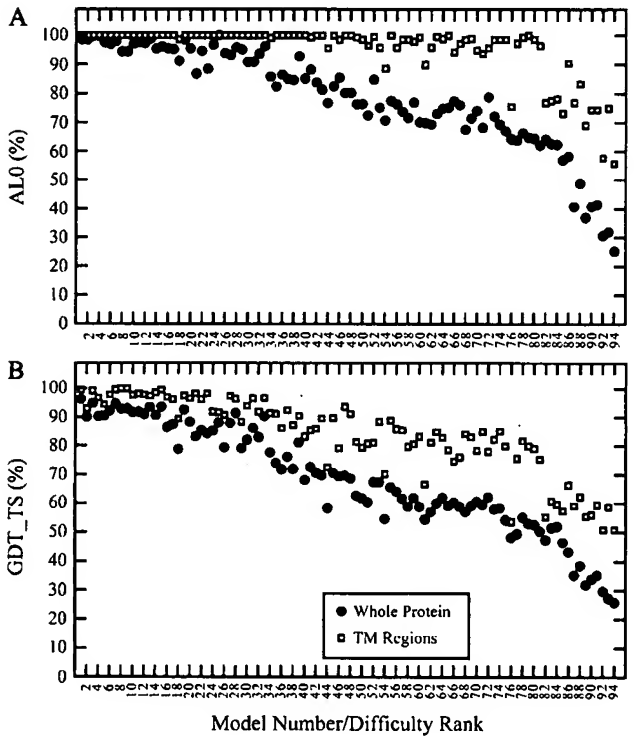


FIGURE 2 Relationship between model quality and model-building difficulty. (*Top*) Alignment accuracy measured by AL0 for the whole protein (●) and transmembrane regions (□). (*Bottom*) Structural accuracy measured by GDT_TS for the whole protein (●) and transmembrane regions (□).

The AL0 scores of the transmembrane regions approach 100% in the majority of the models, whereas the GDT_TS scores for the same regions are often below 100%, suggesting that the errors in the easier models are local deviations that might be removed given an effective refinement protocol.

Secondary-structure prediction accuracy

We ran three different programs on a nonredundant set of membrane proteins of known structure and compared the results with assignments calculated using DSSP (Table 1). The per-residue three-state accuracy (helix, strand, or coil) of the three methods was found to be between 68 and 79%, which is comparable to the ~76% found for globular proteins (25,26,45,46). Similar results were obtained for the composite prediction used by HMAP. Note that the standard deviations are large in all cases, especially for PHDsec and JNET, reflecting a variation in scores that is larger than the 7–10% deviation found for soluble proteins. When considering only the hydrophobic cores, as defined using TMDET, the accuracy improves further, especially for PSIPRED (87%). Comparing the different fold types, we found that α-helical residues in membrane proteins (particularly in the membrane regions) are on average more accurately predicted than β-strand residues, although the data set is smaller for the latter, making such comparisons tentative.

Sequence-based profile alignments

We compare the accuracy of membrane-protein sequence alignments and the models based thereon using the methodologies described in the Methods section. Comparing the two ClustalW methods using the AL0 scores of the respective models (Fig. 3 and Table 2), the sequence-to-profile alignments are more accurate than sequence-to-sequence alignments at low sequence identities. This is in line with results for nonmembrane proteins (47,48). However, in the range of 40–50% sequence identity, the sequence-to-profile alignments are less accurate than the sequence-to-sequence alignments. This has previously been observed for globular protein alignments with ClustalW (28,49).

We also compare the ClustalW alignment results with those of other recently developed multiple-sequence alignment algorithms, namely, Probcons, T-Coffee, and Muscle, which have been reported to be more accurate than ClustalW for globular protein sequence alignments (31–33). Not all of these methods were able to align single sequences to a sequence profile; thus, for each method, we generated multiple-sequence alignments using the PSI-BLAST hits for both query and template (see Methods). The ClustalW multiple-sequence alignments were more accurate than the sequence-to-profile alignments, based on the AL0 scores of the corresponding models (Table 2). Comparing ClustalW multiple-sequence alignments with those of other methods

TABLE 1 Secondary-structure prediction accuracy

	Residues*	PSIPRED	PHDsec	JNET	Composite†
Whole					
All	19,540	79.2 (10.9)	67.6 (17.1)	69.2 (17.6)	77.6 (12.6)
α	15,350	80.0 (10.4)	67.5 (17.6)	69.7 (18.3)	78.5 (12.4)
β	4190	74.1 (13.0)	68.1 (13.4)	65.6 (12.0)	71.8 (12.9)
TMDET					
All	5441	87.3 (16.7)	65.2 (30.2)	71.1 (27.9)	82.3 (20.3)
α	4386	89.6 (13.9)	65.9 (31.7)	73.6 (28.8)	84.7 (19.1)
β	1055	72.2 (24.5)	61.1 (18.8)	54.4 (11.2)	66.5 (21.3)

Average (and standard deviation) of the three-state accuracy, Q_3, for several secondary-structure prediction methods. Q_3 is measured as the percentage of residues that are correctly predicted as helix, strand, or coil relative to the DSSP assignment. Results were averaged either over the whole structure or over the hydrophobic regions as defined by TMDET and separated into helix bundles (α) and β-barrels (β).
*Number of residues in each subset.
†Composite prediction used by HMAP.

using the signed rank test, the newer methods appear to offer significant improvement over ClustalW. Closer inspection reveals that this difference is due to alignments at sequence identities around 40%.

Structure-based profile-profile alignments

The use of the HMAP [1,2] structure-based profile-to-profile alignment method improves the AL0 scores of the models compared with the ClustalW sequence-to-profile alignments and multiple-sequence alignments (Fig. 3 and Table 2). However, the improvement is less obvious when comparing against the newer multiple-sequence alignment methods and in particular with T-Coffee. The most significant improvement in AL0 obtained from HMAP is seen for the most difficult alignments, with sequence identities of <10%. HMAP [1,2,3] alignments are better than the HMAP [1,2] alignments, especially for pairs of sequences with identities of 0–30%. Three-dimensional information is incorporated here using structural alignment of the available templates to guide the combination of their sequence information, as well as the assignment of weights to the core regions (see Methods). Clearly the higher precision achieved by combining template information in this way leads to greater accuracy in the alignments.

In summary, the HMAP [1,2] and HMAP [1,2,3] structure-based profile-to-profile alignments result in the most accurate models of all the methods compared here. However, the alignments obtained from HMAP are not optimal as defined by the structure-based alignments, which obviously limits the accuracy of the models built on these alignments.

Bipartite alignments

All the alignments presented so far, whether sequence- or profile-based, were calculated using the BLOSUM62 amino acid substitution matrix, which was developed for globular proteins (50). It has been suggested that bipartite alignments, which use different substitution matrices for the transmembrane and water-soluble regions, might be more appropriate for membrane proteins (10,16). We tested the effect of using a bipartite approach in a sequence-to-sequence alignment scheme (10,16) on the HOMEP data set using a simple dynamic programming algorithm where the PHAT matrix (16) was applied to the known transmembrane regions in the template and the BLOSUM62 substitution matrix was used for the remaining residues. Note that in contrast to the STAM method (17), we do not align the transmembrane segments separately and then add the loop regions, but rather align the whole sequence and choose the substitution matrix depending on the assignment of each position (10,16). The bipartite alignments result in models with lower AL0 scores than when BLOSUM62 is used throughout (Fig. 4 and Table 3); similar results are observed using Q scores. Using the TM

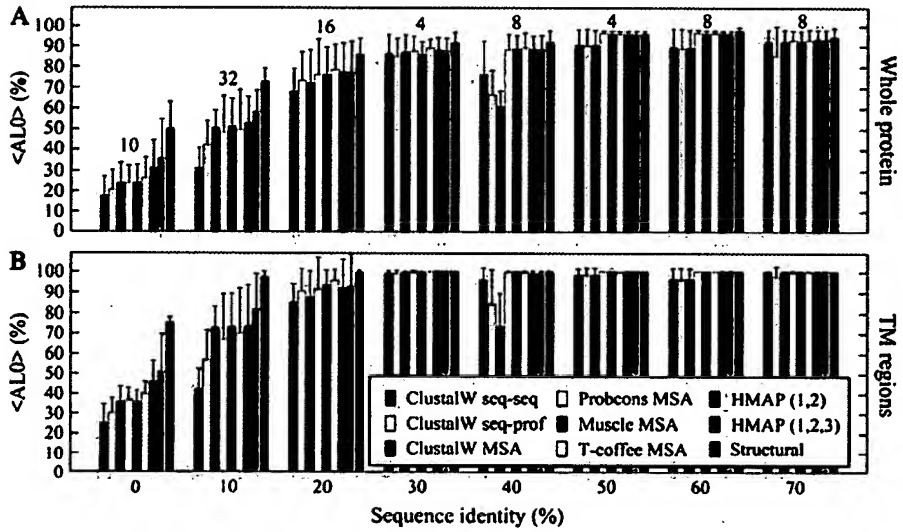


FIGURE 3 Accuracy of membrane protein sequence alignments/homology models obtained from different sequence alignment methods as a function of sequence identity. Results are given for (*A*) the whole protein and (*B*) the transmembrane regions. The average AL0 score is given over all alignments/models within a window of 10% sequence identity, and error bars indicate the standard deviation over that window. Numbers correspond to the number of alignments in each window and apply to both plots. Abbreviations: seq-seq, sequence-to-sequence alignment; seq-profile, sequence-to-profile alignment; and MSA, multiple sequence alignment. The two HMAP labels indicate profile-to-profile alignments.

TABLE 2 The number of HOMEP alignments out of 90 for which a method gives a higher score for the whole/transmembrane regions

AL0	CW seq-seq	CW seq-prof	CW MSA	Probcons	Muscle	T-coffee	HMAP [1,2]	HMAP [1,2,3]
CW seq-seq	–	56/53	58/56	77/64	78/71	79/67	77/64	80/69
CW seq-prof	33/24	–	59/54	73/57	72/67	75/64	77/63	80/67
CW MSA	19/13	28/20	–	57/46	61/58	65/53	69/55	73/61
Probcons	10/9	17/18	29/26	–	32/29	56/38	52/34	55/39
Muscle	10/5	15/12	25/18	55/31	–	60/43	52/36	58/44
T-Coffee	10/7	14/12	23/20	32/15	23/17	–	41/31	44/36
HMAP[1,2]	12/11	13/14	20/17	36/23	35/27	40/26	–	34/32
HMAP[1,2,3]	7/6	9/10	16/11	32/18	29/21	35/21	15/12	–
Q	CW seq-seq	CW seq-prof	CW MSA	Probcons	Muscle	T-coffee	HMAP [1,2]	HMAP [1,2,3]
CW seq-seq	–	49/53	50/54	67/60	65/64	63/64	68/57	71/62
CW seq-prof	36/18	–	57/50	72/58	67/59	74/59	76/58	78/65
CW MSA	28/14	29/19	–	56/44	57/48	59/48	64/52	70/55
Probcons	19/10	18/14	30/22	–	32/34	47/34	54/36	58/41
Muscle	22/9	21/11	28/22	51/31	–	55/43	62/43	69/51
T-Coffee	17/2	16/10	23/18	35/28	26/22	–	44/32	55/41
HMAP[1,2]	13/10	10/13	20/15	29/22	25/20	39/26	–	39/35
HMAP[1,2,3]	11/7	8/9	14/13	27/20	17/17	28/19	11/6	–

Number of times that the alignments from the method in a given column have higher scores than the method in the corresponding row for whole protein/transmembrane regions, using the AL0 score and the *Q*-score. The total number of query-template pairs used was 90, i.e., excluding alignments with 110v. Abbreviations: CW, ClustalW; seq-seq, sequence-to-sequence; seq-prof, sequence-to-profile; MSA, multiple-sequence alignment. For example, the upper right cell in a) reads as follows: The HMAP [1,2,3] alignments give better scores than ClustalW sequence-to-sequence alignments 80 times using the AL0 score for the whole protein, and 69 times for the transmembrane regions only. When only the transmembrane regions are considered, two methods are more likely to give exactly the same result than when the whole sequence is considered since these regions are less variable, and thus the differences tend to be smaller in the former case.

definition of the transmembrane region (see Methods), the bipartite alignments were worse still, which reflects the unsuitability of the PHAT matrix for residues in the bilayer interfacial region.

Since PHAT was developed using transmembrane helices and not β-strands, we also separated the results by fold type (Table 3). As expected, the bipartite scheme worsens the alignments of the β-barrels, whereas the alignments of the helical bundles are very similar to when BLOSUM62 alone is used. Overall, in the most basic bipartite implementation, the PHAT substitution matrix does not appear to improve sequence-to-sequence alignments of membrane proteins.

Errors in individual alignments

For a few models we observe that the alignments generated using either HMAP [1,2] or HMAP [1,2,3] profiles were less accurate than the ClustalW sequence-to-profile alignments. The largest differences are found for the TonB-coupled receptor family, most strikingly in the models where BtuB (PDB code: 1nqe) is the query or where FepA (PDB code: 1fep) is the query. These errors are likely caused by the low secondary-structure prediction accuracy for the long β-strands in the TonB-coupled receptor family, which is 65.1% with PSIPRED. Other poor quality alignments are found for the seven transmembrane helix models (see Opsins in Supplementary Material Table 1), when rhodopsin (PDB code: 1u19) is either the query or the template, although the HMAP alignments are usually better than the ClustalW sequence-to-profile alignments. The structure of bovine rhodopsin is significantly different from that of the three bacterial opsins: the transmembrane helices of rhodopsin are more distorted and it contains an additional (interfacial) helix, a small β-sheet, and much longer loops and termini. These differences, along with extremely low sequence identities, combine to yield relatively poor quality alignments and models for this family.

DISCUSSION

Membrane protein homology modeling benchmark

In this study, we have presented a detailed analysis of the applicability of sequence alignment and homology modeling methods to integral membrane proteins. The HOMEP data set is key to the analysis, since it covers a range of fold types and sequence identities and thus provides a comprehensive benchmark of realistic modeling situations. Using this benchmark we show that similar trends exist with respect to the sequence-structure relationship (43,44) and to alignment accuracy (28) as are observed for water-soluble proteins. In addition, with this benchmark, it is possible to predict the likely accuracy of a homology model, assuming that an accurate alignment can be achieved and that no refinement is attempted. We find that the relationship between sequence identity and structure similarity is similar to that observed for water-soluble proteins, so that experience based on model accuracy for soluble proteins should be applicable to

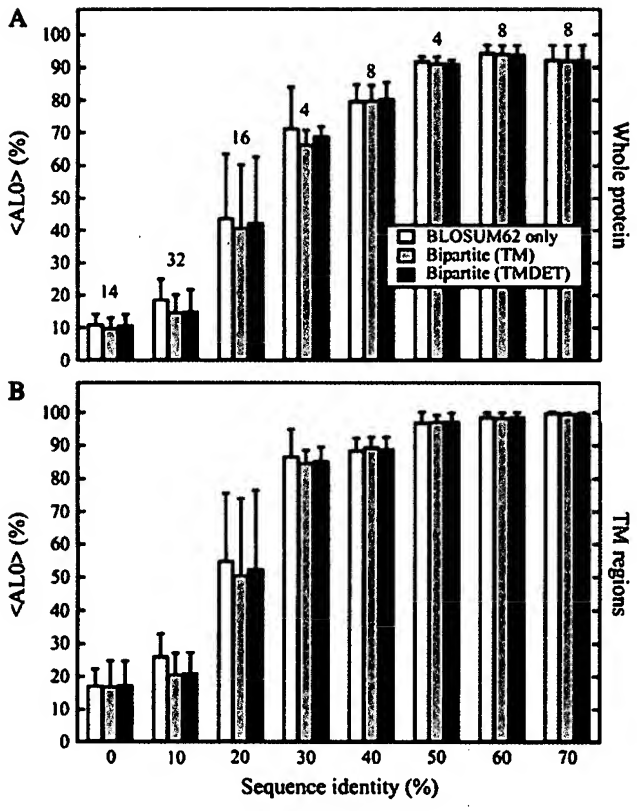


FIGURE 4 Accuracy of bipartite sequence-to-sequence alignments of membrane proteins obtained with different substitution matrices. See legend to Fig. 3 for more details.

membrane proteins as well. For the transmembrane regions the expected model accuracy is higher than for the whole protein. For example, at 50% sequence identity to the template, a model is expected to have a Cα-RMSD of ~1 Å from the native structure (~95% GDT_TS) in the transmembrane regions. Indeed, an acceptable model of, say, 2 Å Cα-RMSD in the transmembrane regions (~85% GDT_TS) is possible for most proteins above 30% sequence identity. In contrast, below ~25% sequence identity, which is the similarity

TABLE 3 Signed rank test using AL0 values for BLOSUM62-only against bipartite alignments for the whole/transmembrane regions

Transmembrane definition*	Better†	Worse‡	Total§
TMDET	57/53	23/21	94
TM	63/54	19/26	94
TMDET: helix bundles	19/17	17/19	46
TMDET: β-barrels	38/36	6/2	48
TM: helix bundles	27/21	13/18	46
TM: β-barrels	36/33	6/8	48

*Definition of residues treated as transmembrane in the bipartite scheme (see Methods).
†Number of times that BLOSUM62-only alignments are better.
‡Number of times that BLOSUM62-only alignments are worse.
§Number of alignments tested.

of many G-protein-coupled receptors to bovine rhodopsin—the only available template—a model may have a transmembrane Cα-RMSD from the native above 3.0 Å (~75% GDT_TS). The accuracy of the complete model, including all extramembranous regions, will be expected to be lower than that of the transmembrane region alone.

This analysis indicates the accuracy of a model assuming that the conformation of the template structure reflects the desired conformation of the query protein. However, many membrane proteins are believed to undergo conformational changes during functional processes. Homology models cannot be expected to accurately predict such conformational changes per se: only the conformation closest to that of the chosen template will be adequately represented. Thus, the accurate prediction of many different functional conformations of a membrane protein will require template structures in equivalent conformations to be solved.

Membrane protein sequence alignments

Our analysis of sequence alignment algorithms indicates that those methods that have proved effective for water-soluble proteins work for membrane proteins as well. There is a clear progression in alignment accuracy when recently developed multiple sequence alignment (MSA) algorithms are used and additional improvements are obtained with HMAP's profile-to-profile alignment algorithm. Moreover, the increased use of structural information in the HMAP [1,2,3] alignments yields improvements relative to the HMAP [1,2] alignments. We note that ClustalW (29) is widely used to create sequence alignments for membrane proteins (51–56). Our results suggest that future work would benefit from the use of profile-to-profile methods and/or more advanced MSA techniques.

Our results on a simple bipartite sequence-to-sequence alignment method using the membrane-protein-specific substitution matrix PHAT show no significant improvement in the alignment quality over a traditional alignment using BLOSUM62. Originally, PHAT was shown to improve sensitivity in sequence database searches of membrane proteins (16). However, since database searching aims to best discriminate between similar and dissimilar proteins, rather than to achieve the correct global alignment of two sequences, the optimal parameters for the two applications may differ. There have also been some reported improvements in alignment accuracy using PHAT within the program STAM (17), which might be attributable to the separation and independent alignment of the transmembrane and nontransmembrane regions and to differences in gap penalties, rather than to the choice of substitution matrix. Clearly, the usefulness of membrane-protein-specific substitution matrices is dependent on the context, suggesting that the contribution of the choice of matrix should be carefully assessed in future applications.

Many other strategies have been presented for the alignment of membrane protein sequences (17,57–59) and for

database searches (60,61). For example, probable transmembrane regions and loop regions have been aligned separately as independent segments (17,58) and then reassembled. Alignment of hydropathy profiles, rather than of primary sequences, has also been proposed (57). These methods have not been assessed here, either because they are not automated or because they were only suitable for helical proteins. However, it would be interesting to see how these methods compare with the profile-to-profile methods in terms of membrane protein alignment accuracy. Indeed, comparison of models from fully automated methods with those generated by experts in the field (with manual adjustment of alignments, for example) suggests that the manual approaches can lead to higher model accuracies (62). This has relevance to the alignments used in, e.g., G-protein-coupled receptor modeling (63), which have often required manual intervention. Nevertheless, a poor initial alignment may introduce errors that are missed during manual adjustment, particularly at low sequence identities, emphasizing the importance of accurate alignment algorithms.

Secondary-structure prediction

The success of the profile-to-profile methods is dependent on the accurate prediction of secondary structures in the query protein. We have shown that current secondary-structure prediction algorithms, and in particular PSIPRED, are only slightly less accurate for membrane proteins than they are for water-soluble proteins. This is rather surprising, since amino acids in membranes are reported to have different secondary-structure propensities (13–15) and because early prediction methods (64) gave results in poor agreement with experimental data for membrane proteins (65). Our results, which instead assess more recent, neural-network-based approaches using a larger set of high-resolution data, are supported by a previous study of membrane protein β-barrel prediction (66) in which similar results were obtained using PSIPRED (73%). (To our knowledge, no similar study has previously been attempted for helical membrane proteins.)

Neural networks derived from soluble proteins might have been expected to perform poorly on membrane proteins for two reasons: the membrane region imposes different secondary-structure propensities on amino acids, and the algorithms were not trained on membrane protein structures. Their success for membrane proteins may be due to the detection of the periodicity that is present in both sets of proteins. Even though the periodicity is effectively inverted, i.e., the surface of transmembrane regions is more hydrophobic than the interior whereas the surface of water-soluble proteins is more hydrophilic than the interior, the existence of a regular periodic pattern alone may be sufficient to obtain good prediction accuracy. In membrane protein β-barrels, the strands often extend far beyond the hydrophobic bilayer core where their properties are likely to strongly resemble the alternating patterns of water-soluble protein β-strands. However, the five to seven residues that comprise the membrane-spanning part of the strands may have a more complex pattern: the outer face of the barrel will be predominantly hydrophobic, whereas the interior face properties will depend on whether the barrel is filled with protein or water. This might explain the lower accuracy seen for the predictions on the hydrophobic TMDET regions of the β-barrels compared with the whole structures, although definitive interpretations are difficult due to the small number of structures (Table 1).

Secondary structure versus transmembrane prediction

Since they do not predict the same property, it is somewhat specious to directly compare the accuracies of secondary-structure predictions with those of transmembrane predictions. For reference, however, we note that the best-performing transmembrane-helix predictors have two-state per-residue accuracies (i.e., whether a residue is in the membrane or not) of ~80% (67,68). Their accuracy at the segment level (i.e., whether a membrane-spanning helix is detected or not) is generally higher, between 85 and 99%. In the case of the β-barrel predictors, per-residue accuracies of ~82% have been achieved (69). Thus, both the transmembrane helix and transmembrane strand methods are only slightly more accurate than the secondary-structure prediction algorithms. It is noteworthy, though, that as a consequence of the low number of structures available, accuracies for transmembrane predictions may be inflated by overtraining or by tests using proteins that were also included within the training set (68). In contrast, the secondary-structure prediction algorithms were solely trained on water-soluble proteins.

CONCLUSIONS

Using the HOMEP data set, we show that the construction of membrane protein homology models follows similar general rules to the construction of water-soluble models. That is, the expected accuracy of a membrane protein model will be similar to that of a water-soluble protein, assuming that similar alignment accuracy can be achieved. However, as a result of the low numbers of experimental structures of membrane proteins currently available, many candidate proteins for modeling are likely to have low sequence identities to their templates, so that accurate alignment of their sequences will be especially challenging. Our results suggest that more accurate alignments for such proteins can be achieved using structure-based profile alignment methods that have been developed for water-soluble proteins. In the future, however, it may be possible to incorporate information specific to membrane proteins—such as the location of hydrophobic transmembrane regions—within these methods to make alignments and homology models of membrane proteins even more accurate.

SUPPLEMENTARY MATERIAL

An online supplement to this article can be found by visiting BJ Online at http://www.biophysj.org.

We thank Shoshana Posy, Donald Petrey, Henry Bigelow, and Andrew Kernytsky for helpful discussions, and José Faraldo-Gómez for useful comments on the manuscript.

This work was supported by the National Science Foundation under grant No. MCB-0416708.

REFERENCES

1. Jones, D. T. 1998. Do transmembrane protein superfolds exist? *FEBS Lett.* 423:281–285.
2. Wallin, E., and G. von Heijne. 1998. Genome-wide analysis of integral membrane proteins from eubacterial, archean, and eukaryotic organisms. *Protein Sci.* 7:1029–1038.
3. Krogh, A., B. Larsson, G. von Heijne, and E. L. L. Sonnhammer. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* 305:567–580.
4. Drews, J. 2000. Drug discovery: a historical perspective. *Science.* 287: 1960–1964.
5. Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. 2000. The protein data bank. *Nucleic Acids Res.* 28:235–242.
6. Petrey, D., and B. Honig. 2005. Protein structure prediction: inroads to biology. *Mol. Cell.* 20:811–819.
7. Wallin, E., T. Tsukihara, S. Yoshikawa, G. von Heijne, and A. Elofsson. 1997. Architecture of helix bundle membrane proteins: an analysis of cytochrome *c* oxidase from bovine mitochondria. *Protein Sci.* 6:808–815.
8. Liu, Y., D. M. Engelman, and M. Gerstein. 2002. Genomic analysis of membrane protein families: abundance and conserved motifs. *Genome Biol.* 3:research0054.0051–0054.0012.
9. Donnelly, D., J. P. Overington, S. V. Ruffle, J. H. A. Nugent, and T. L. Blundell. 1993. Modelling α-helical transmembrane domains: the calculation and use of substitution tables for lipid-facing residues. *Protein Sci.* 2:55–70.
10. Jones, D. T., W. R. Taylor, and J. M. Thornton. 1994. A mutation data matrix for transmembrane proteins. *FEBS Lett.* 339:269–275.
11. Rees, D. C., L. DeAntonio, and D. Eisenberg. 1989. Hydrophobic organization of membrane proteins. *Science.* 245:510–513.
12. Eyre, T. A., L. Partridge, and J. M. Thornton. 2004. Computational analysis of alpha-helical membrane protein structure: implications for the prediction of 3D structural models. *Protein Eng. Des. Sel.* 17: 613–624.
13. Li, S.-C., and C. M. Deber. 1994. A measure of helical propensity for amino acids in membrane environments. *Nat. Struct. Biol.* 1:368–373.
14. Blondelle, S. E., B. Forood, R. A. Houghten, and E. Pérez-Payá. 1997. Secondary structure induction in aqueous vs membrane-like environments. *Biopolymers.* 42:489–498.
15. Monné, M., I. Nilsson, A. Elofsson, and G. von Heijne. 1999. Turns in transmembrane helices: determination of the minimal length of a "helical hairpin" and derivation of a fine-grained turn propensity scale. *J. Mol. Biol.* 293:807–814.
16. Ng, P. C., J. G. Henikoff, and S. Henikoff. 2000. PHAT: a transmembrane-specific substitution matrix. *Bioinformatics.* 16:760–766.
17. Shafrir, Y., and H. R. Guy. 2004. STAM: simple transmembrane alignment method. *Bioinformatics.* 20:758–769.
18. Granseth, E., G. von Heijne, and A. Elofsson. 2005. A study of the membrane-water interface region of membrane proteins. *J. Mol. Biol.* 346:377–385.
19. Riek, R. P., I. Rigoutsos, J. Novotny, and R. M. Graham. 2001. Non-α-helical elements modulate polytopic membrane protein architecture. *J. Mol. Biol.* 306:349–362.
20. Bowie, J. U. 2005. Solving the membrane protein folding problem. *Nature.* 438:581–589.
21. Thompson, J. D., P. Koehl, R. Ripp, and O. Poch. 2005. BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. *Proteins.* 61:127–136.
22. Kabsch, W., and C. Sander. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers.* 22:2577–2637.
23. Tusnady, G. E., Z. Dosztanyi, and I. Simon. 2005. TMDET: web server for detecting transmembrane regions of proteins by using their 3D coordinates. *Bioinformatics.* 21:1276–1277.
24. Tusnady, G. E., Z. Dosztanyi, and I. Simon. 2005. PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank. *Nucleic Acids Res.* 33:D275–D278.
25. Jones, D. T. 1999. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* 292:195–202.
26. Cuff, J. A., and G. J. Barton. 1999. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins.* 40:502–511.
27. Rost, B. 1996. PHD: predicting 1D protein structure by profile-based neural networks. *Methods Enzymol.* 266:525–539.
28. Tang, C. L., L. Xie, I. Y. Y. Koh, S. Posy, E. Alexov, and B. Honig. 2003. On the role of structural information in remote homology detection and sequence alignment methods using hybrid sequence profiles. *J. Mol. Biol.* 334:1043–1062.
29. Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL_W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22:4673–4680.
30. Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403–410.
31. Notredame, C., D. G. Higgins, and J. Heringa. 2000. T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* 302:205–217.
32. Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–1797.
33. Do, C. B., M. S. P. Mahabhashyam, M. Brudno, and S. Batzoglou. 2005. ProbCons: probabilistic consistency-based multiple sequence alignment. *Genome Res.* 15:330–340.
34. Ohlson, T., B. Wallner, and A. Elofsson. 2004. Profile-profile methods provide improved fold recognition: a study of different profile-profile alignment methods. *Proteins.* 57:188–197.
35. Yang, A. S., and B. Honig. 2000. An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance. *J. Mol. Biol.* 301:665–678.
36. Petrey, D., A. Nicholls, and B. Honig. 2003. GRASP2: visualization, surface properties and electrostatics of macromolecular structures and sequences. *Methods Enzymol.* 374:492–509.
37. Sali, A., and T. L. Blundell. 1993. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* 234:779–815.
38. Moult, J., K. Fidelis, A. Tramontano, B. Rost, and T. Hubbard. 2005. Critical assessment of methods of protein structure prediction (CASP)—round 6. *Proteins.* 61:3–7.
39. Zemla, A. 2003. LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res.* 31:3370–3374.
40. Moult, J., K. Fidelis, A. Zemla, and R. E. Hubbard. 2001. Critical assessment of methods of protein structure prediction (CASP): round IV. *Proteins.* 45:2–7.
41. Venclovas, C., A. Zemla, K. Fidelis, and J. Moult. 2003. Assessment of progress over the CASP experiments. *Proteins.* 53:585–595.

42. Wilson, C. A., J. Kreychman, and M. Gerstein. 2000. Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J. Mol. Biol.* 297:233–249.

43. Flores, T. P., C. A. Orengo, D. S. Moss, and J. M. Thornton. 1993. Comparison of conformational characteristics in structurally similar protein pairs. *Protein Sci.* 2:1811–1826.

44. Chothia, C., and A. M. Lesk. 1986. The relation between the divergence of sequence and structure in proteins. *EMBO J.* 5:823–826.

45. Rost, B., and V. A. Eyrich. 2001. EVA: large-scale analysis of secondary structure prediction. *Proteins.* 45:192–199.

46. Rost, B. 2001. Review: protein secondary structure prediction continues to rise. *J. Struct. Biol.* 134:204–218.

47. Thompson, J. D., F. Plewniak, and O. Poch. 1999. A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res.* 27:2682–2690.

48. Wallace, I. M., G. Blackshields, and D. G. Higgins. 2005. Multiple sequence alignments. *Curr. Opin. Struct. Biol.* 15:261–266.

49. Elofsson, A. 2002. A study on protein sequence alignment quality. *Proteins.* 46:330–339.

50. Henikoff, S., and J. G. Henikoff. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA.* 89:10915–10919.

51. Ogawa, H., and C. Toyoshima. 2002. Homology modeling of the cation binding sites of Na^+K^+-ATPase. *Proc. Natl. Acad. Sci. USA.* 99:15977–15982.

52. Casadio, R., I. Jacoboni, A. Messina, and V. De Pinto. 2002. A 3D model of the voltage-dependent anion channel (VDAC). *FEBS Lett.* 520:1–7.

53. Yang, Q., X. Wang, L. Ye, M. Mentrikoski, E. Mohammadi, Y.-M. Kim, and P. C. Maloney. 2005. Experimental tests of a homology model for OxlT, the oxalate transporter of *Oxalobacter formigenes*. *Proc. Natl. Acad. Sci. USA.* 102:8513–8518.

54. Kuhlbrandt, W., J. Zeelen, and J. Dietrich. 2002. Structure, mechanism, and regulation of the *Neurospora* plasma membrane H^+-ATPase. *Science.* 297:1692–1696.

55. Bostina, M., B. Mohsin, W. Kühlbrandt, and I. Collinson. 2005. Atomic model of the *E. coli* membrane-bound protein translocation complex SecYEG. *J. Mol. Biol.* 352:1035–1043.

56. Oyedotun, K. S., and B. D. Lemire. 2004. The quaternary structure of the *Saccharomyces cerevisiae* succinate dehydrogenase: homology modeling, cofactor docking and molecular dynamics simulation studies. *J. Biol. Chem.* 279:9424–9431.

57. Lolkema, J. S., and D. J. Slotboom. 1998. Estimation of structural similarity of membrane proteins by hydropathy profile alignment. *Mol. Membr. Biol.* 15:33–42.

58. Bissantz, C., A. Logean, and D. Rognan. 2004. High-throughput modeling of human G-protein coupled receptors: amino acid sequence alignment, three-dimensional model building, and receptor library screening. *J. Chem. Inf. Comput. Sci.* 44:1162–1176.

59. Cserzo, M., J.-M. Bernassau, I. Simon, and B. Maigret. 1994. New alignment strategy for transmembrane proteins. *J. Mol. Biol.* 243: 388–396.

60. Clements, J. D., and R. E. Martin. 2002. Identification of novel membrane proteins by searching for patterns in hydropathy profiles. *Eur. J. Biochem.* 269:2101–2107.

61. Hedman, M., H. Deloof, G. Von Heijne, and A. Elofsson. 2002. Improved detection of homologous membrane proteins by inclusion of information from topology predictions. *Protein Sci.* 11: 652–658.

62. Tress, M. L., I. Ezkurdia, O. Graña, G. López, and A. Valencia. 2005. Assessment of predictions submitted for the CASP6 comparative modeling category. *Proteins.* 61:27–45.

63. Fanelli, F., and P. G. De Benedetti. 2005. Computational modeling approaches to structure-function analysis of G protein-coupled receptors. *Chem. Rev.* 105:3297–3351.

64. Chou, P. Y., and G. D. Fasman. 1974. Conformational parameters for amino acids in helical, β-sheet and random coil regions calculated from proteins. *Biochemistry.* 13:211–222.

65. Wallace, B. A., M. Cascio, and D. L. Mielke. 1986. Evaluation of methods for the prediction of membrane protein secondary structures. *Proc. Natl. Acad. Sci. USA.* 83:9423–9427.

66. Bagos, P. G., T. D. Liakopoulos, I. C. Spyropoulos, and S. J. Hamodrakas. 2004. PRED-TMBB: a web server for predicting the topology of β-barrel outer membrane proteins. *Nucleic Acids Res.* 32:W400–W404.

67. Chen, C. P., and B. Rost. 2002. State-of-the-art in membrane protein prediction. *Appl. Bioinformatics.* 1:21–35.

68. Chen, C. P., A. Kernytsky, and B. Rost. 2002. Transmembrane helix predictions revisited. *Protein Sci.* 11:2774–2791.

69. Bagos, P. G., T. Liakopoulos, and S. Hamodrakas. 2005. Evaluation of methods for predicting the topology of beta-barrel outer membrane proteins and a consensus prediction method. *BMC Bioinformatics.* 6:7.

doi:10.1016/j.jmb.2006.01.023 J. Mol. Biol. (2006) 357, 1566–1574

JMB

Available online at www.sciencedirect.com
SCIENCE DIRECT®

ELSEVIER

Antibody Elbow Angles are Influenced by their Light Chain Class

Robyn L. Stanfield[1]*, Adam Zemla[2], Ian A. Wilson[1,3] and Bernhard Rupp[2,4]*

[1]*Department of Molecular Biology, The Scripps Research Institute, 10550 N. Torrey Pines Road, La Jolla, CA 92037 USA*

[2]*Lawrence Livermore National Laboratory, 7000 East Avenue Livermore, CA 94551, USA*

[3]*Skaggs Institute for Chemical Biology, The Scripps Research Institute, 10550 N. Torrey Pines Road, La Jolla, CA 92037 USA*

[4]*q.e.d. Life Science Discoveries 6487 Half Dome Court Livermore, CA 94551, USA*

**Corresponding authors*

We have examined the elbow angles for 365 different Fab fragments, and observe that Fabs with λ light chains have adopted a wider range of elbow angles than their κ chain counterparts, and that the λ light chain Fabs are frequently found with very large (>195°) elbow angles. This apparent hyperflexibility of λ chain Fabs may be due to an insertion in their switch region, which is one residue longer than in κ chains, with glycine occurring most frequently at the insertion position. A new, web-based computer program that was used to calculate the Fab elbow angles is described.

Keywords: antibody; elbow angle; lambda; kappa; computer program

Introduction

Antibodies are composed of two light (L; ~25,000 Da) and two heavy (H; ~50,000 Da) polypeptide chains that combine to form one Fc and two Fab modules that can be isolated as functional fragments by proteolytic cleavage of the intact immunoglobulin. Within each Fab fragment are two types of distinct structural domains termed variable (V_L,V_H) and constant (C_L,C_H1), with the amino acid residues linking V_L to C_L and V_H to C_H1 called switch residues. Since the early 1970s, when Fab and light-chain dimer structures first became available, it was noted that these fragments displayed a variability in the angle between their variable and constant domains,[1] referred to as the elbow bend or elbow angle and defined as the angle between the pseudo-2-fold axes relating V_L to V_H, and C_L to C_H1 (Figure 1). While these early antibody structures sparked speculation that the elbow angle might change in response to ligand binding,[2] no convincing data have since been found to support this theory. Rather, it appears that the elbow angle may simply serve to increase Fab flexibility, thus enhancing the ability of an antibody to bind bivalently to ligands arranged on a pathogen surface, such as a virus or bacteria.

The Fab elbow angle (Figure 1) is a useful descriptor of the overall topology of the Fab fragment, serving as a measure for the relative disposition of the variable *versus* the constant domains. The elbow angle is almost always cited in Fab structure reports that include comparison of liganded *versus* unliganded Fab structures, and assessment of Fab switch region flexibility. In order to simplify the elbow angle calculation, we have developed a web-based program to more readily calculate the elbow angle for RCSB Protein DataBank (PDB) formatted Fab coordinates. We have tested this method by calculating elbow angles for 536 Fab fragments from the PDB (of which 365 are non-redundant). The elbow angle calculations are consistent with previous compilations but now clearly demonstrate that the propensity for λ light chains to assume elbow angles beyond 195° is significant compared to κ light chains.

Abbreviations used: V_L, variable light; V_H, variable heavy; C_L, constant light; C_H1, constant heavy; POB, protein Data Bank.

E-mail addresses of the corresponding authors: robyn@scripps.edu; bernhardrupp@sbcglobal.net

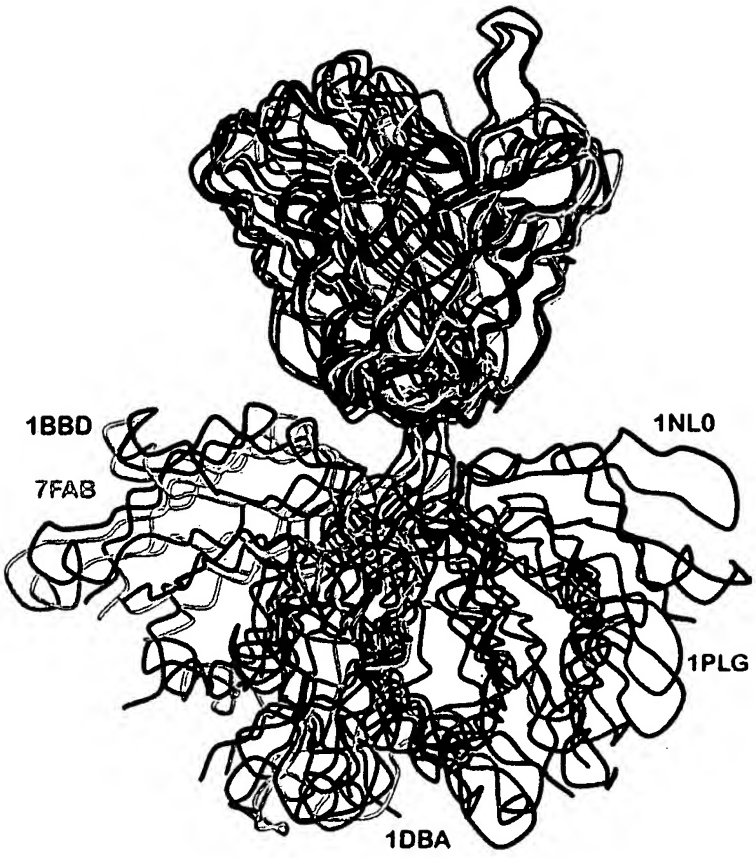


Figure 1. Superposition of a variety of Fabs with different elbow angles. Fabs from PDB files 1bbd (red), 7fab (yellow), 1dba (green), 1plg (cyan), and 1nl0 (blue) have been superimposed on their variable light chain regions. The range of elbow angles is shown from small (1bbd, 127°; 7fab, 132°) to around 180° (1dba, 183°) to large (1plg, 190°; 1nl0, 220°).

Procedures

Calculation of elbow angles

The elbow angle of a Fab antibody fragment is defined as the angle between the two, not necessarily intersecting, pseudo-dyad axes relating the light (V_L) and heavy (V_H) chain variable domains, and the light (C_L) and heavy (C_H1) chain constant domains. Small deviations in the exact locations of the pseudo-dyad axes arise depending on which residues are used for their calculation; however, the resultant deviations for the values of the elbow angles are usually limited to only a few degrees. It is standard practice to use only structurally conserved residues from the antibody framework region for this calculation to eliminate errors due to differences in conformations of the complementarity determining regions. If the Fab coordinates used for the calculation have been numbered in a consistent way (such as the Kabat & Wu convention[3]), these structurally equivalent residues are easily defined. The calculation is then easy; however, most of the Fab structures deposited in the RCSB Protein DataBank[4] are not numbered or labeled consistently, making such comparisons more difficult. The program we use for the V_L-V_H and C_L-C_H1 superposition (LGA[5]) also refines the sequence alignment between the domains and, thus the superposition geometry does not depend critically on the numbering system used for the Fab.

However, additional fine points in the elbow angle calculation need to be addressed. The elbow angle is calculated as the dot product of the V_L-V_H and C_L-C_H1 pseudo-dyad angle, and always computes between 0° and 180°. Although one could readily determine the absolute value in mathematical terms through the sign of the determinant of the basis matrix formed by the two pseudo-dyad vectors and their cross-product vector, this approach does not overcome the problem. Due to the reduction of the information to a single scalar angle value, the relative orientation of the axes is lost, and the solutions become degenerate (imaginable as located on a cone) between 90° and 180°. To regain the domain orientation on an absolute scale and to solve the complement ambiguity, one needs to compare each Fab to a "standard" Fab with a defined elbow angle. In this case, we use unliganded Fab 8F5 (PDB code 1bbd) as the standard, with an elbow angle of 127°. The Fab to be tested is first superimposed *via* its variable domain (V_L-V_H) onto the variable domain of the standard Fab. From this V-aligned orientation, the constant domain (C_L-C_H1) of the test Fab is aligned with the constant domain of the standard Fab, yielding the rotation relative to 1bbd. The sum of the standard Fab's elbow angle (127°) plus the θ_3

domain rotation angle is then used to resolve the degeneracy and to determine whether the elbow angle calculated by the dyad dot product is direct or complementary (>180°). This method has been used extensively in previous reviews and compilations.[6]

Limitations and accuracy

The underlying assumption of the definition of the elbow angle is that the superposition of V_L onto V_H, as well as C_L onto C_H1, is predominantly a 2-fold rotation. In this case, the directional cosine matrix resembles a pure 2-fold rotation with relatively small non-diagonal components and the Euler axis is dominated by a correspondingly large principal component. In more than 95% of the Fabs that we examined, the assumption of near-2-fold superposition operations is justified, and the elbow angle concept fully applicable.

If additional rotational components become significant in the domain superpositions, the Euler axes can deviate from the 2-fold to a point where the two axes become non-opposing and the calculated elbow angle is less than 90°. Even in these cases, a complement ($180-\alpha$) can be interpreted as the elbow angle, but such Fabs need to be examined individually to decide whether the elbow angle definition is still meaningful.

Although elbow angles tend to be reported to a precision of one decimal point, the choice of domain limits and superposition procedure places limitations on the absolute accuracy of the elbow angles. The superposition results can vary, depending on how the domain limits are defined, and depend on the alignment procedure used. With default domain limits ($V_L \leq L107 < C_L$ and $V_H \leq H113 < C_H1$, residues in Kabat numbering), we compared the elbow angles obtained using Kabat-renumbered Fab coordinates, and a fixed set of structurally conserved residues for the superposition (using the program OVRLAP[7]) with those computed using our web application based on automated LGA alignments[5] and original numbering from the PDB. The angles computed ($n=167$) using the different methods agreed with a difference in their mean of −0.12° and a standard deviation of no more than 1.1°. It is interesting to note that molecular dynamics simulations of Fab domain movement in solution show periodic hinge bending fluctuations, with a 2–3° root-mean-square deviation (r.m.s.d.) in elbow angle.[8] The presence of such dynamic fluctuations indicates that the reported precision to a tenth of a degree in the elbow angles is, indeed, overly optimistic, whereas the differences in calculated elbow angles based on different domain superposition techniques are well within the range of the dynamic elbow flexibility. It would, therefore, seem reasonable that any difference in elbow angles below 2–3° should not be considered significant.

In addition, the extent to which crystal packing additionally affects or limits the observed elbow angles is uncertain. Several examples exist where two Fabs in the same crystallographic asymmetric unit display significantly different elbow angles, for example 1jnh (27° difference), 1s78 (22°) and 1ots (21°). However, even within the limits of accuracy discussed above, the elbow angles are still useful as measures of hinge flexibility and for classification purposes.

Analysis of elbow angles

Elbow angles for a total of 536 Fab fragments in 416 PDB entries have been calculated using the automated procedure described above. A non-redundant data set was created by omitting repetitions (such as one antibody in complex with many similar haptens), following a procedure used previously.[9] Fabs crystallizing in the same space group with cell constants within 1% and their elbow angles within ±3° were considered equivalent, resulting in a non-redundant set containing 365 unique Fab structures (Supplementary Data, Table S1). The distributions of elbow angles (**Figure 2**) are distinctly different, depending on the Fab light chain type. Most of the elbow angles larger than 190° belong to the group of 33 unique λ chains in the test set (**Table 1**; **Figure 2**). No correlation is apparent with heavy chain type (**Table 1**). The most frequent space groups for Fab crystal structures follow the general distribution analyzed previously for proteins,[9,10] with $P2_12_12_1$ (32%), $P2_1$ (29%), $P1$ (24%) and $C2$ (13%) being the dominant space groups.

Implementation

The program RBOW is implemented on a Linux platform (Apache web server) using a combination of Perl scripts and standard Fortran90 code. The alignment program used is the local-global alignment program LGA,[5] which assures minimal dependence of the results on the Fab numbering convention used. The program allows upload of PDB format coordinate files, selection of heavy and light chain identifiers, and input of domain boundaries, for which reasonable default values are provided. During the calculation as described above, the program checks for format errors and issues warnings at several levels for unexpected or borderline behavior. Warnings include large coordinate r.m.s.d. values on superposition (>3.5 Å), significant deviations from pseudo-2-fold rotation axes, occurrence of parallel superposition axes, and swapped annotation of L and H chains. The superposition files are available for download and visual inspection, if desired. Coordinates may be uploaded for the elbow angle calculation†.

† http://as2ts.llnl.gov/AS2TS/RBOW/

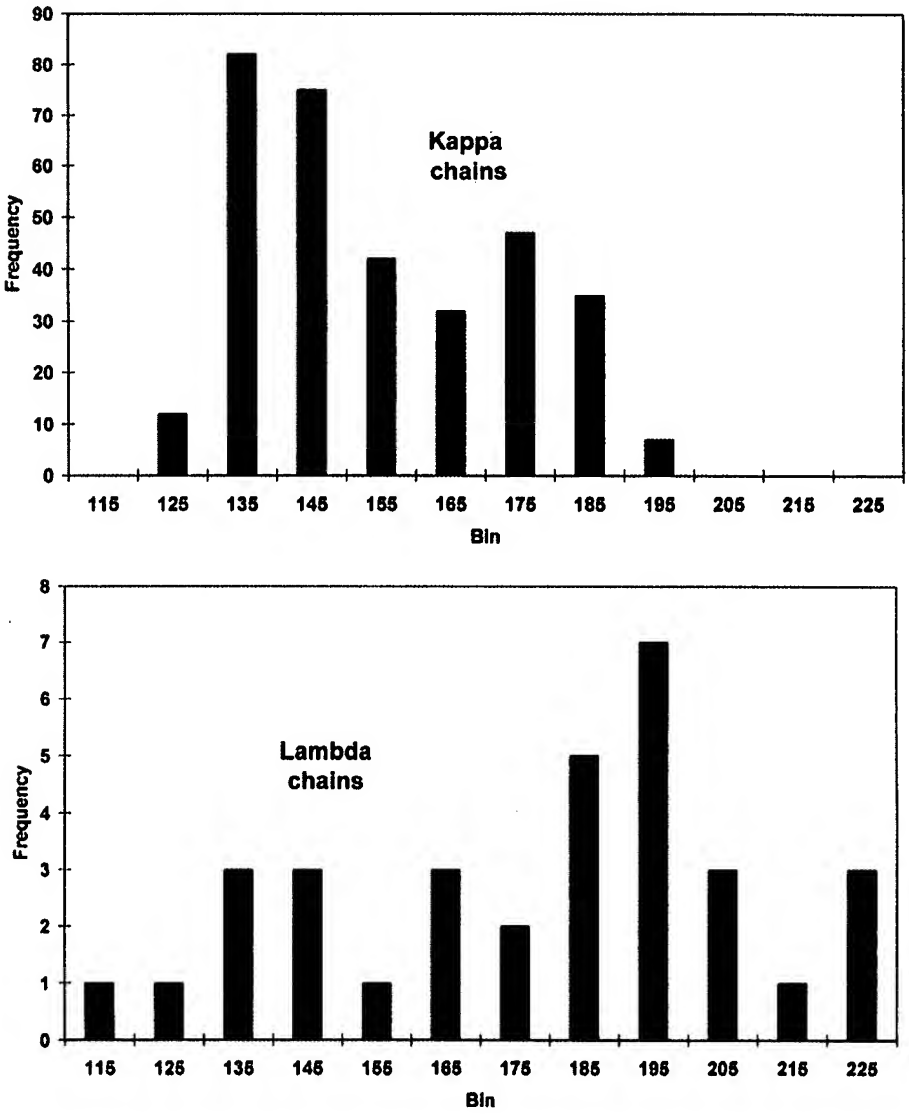


Figure 2. Distribution of elbow angles for κ and λ chain type Fabs. The distinct preference of λ chains to adopt large elbow angles is clearly displayed. Bin numbers indicate the highest value of the corresponding bin.

Table 1. Statistics of non-redundant elbow angle distributions in κ and λ chain Fabs

Antibody light chain type	*N*	<180°	>180°	>195°	Mean	Median	Skew
κ	332	290 (87)	42 (13)	1 (0)	156°	150°	+0.4
λ	33	14 (42)	19 (58)	11 (33)	178°	185°	−0.4
Antibody heavy chain class							
IgG1	239	203 (85)	36 (15)	8 (4)	156°	150°	+0.7
IgG2a	79	63 (79)	16 (21)	0	158°	154°	+0.1
IgG2b	22	19 (86)	3 (14)	0	158°	151°	+0.8

Values in parentheses are percentages. In addition to the moment-based statistics listed above, *F*-tests, two-factor ANOVA (light chain type, heavy chain class) and non-parametric Spearman rank tests confirm that no significant correlations in the data exist beyond the preference of λ type light chains to adopt large elbow angles.

Discussion

The majority of Fab fragment structures in the PDB[4] have κ light chains, with only 37 λ-chain Fab PDB depositions (as of May 2005), half of which have been submitted since the year 2000. This paucity of λ chain Fab structures reflects their lower abundance, particularly in mice where the antibody light chain repertoire is about 5% λ and 95% κ,[11] whereas in humans about 40% of the repertoire is λ. Also, the number of mouse structures far outweighs their human counterparts (of 416 PDB entries, 305 were mouse and 51 human, with the rest chimeric or humanized, rat (4) or hamster (1)). Our results show that of the 12 unique human and 21 unique mouse or hamster Fab structures with λ light chains, about 60% have elbow angles greater than 180°, with 11 instances of elbow angles greater than 195° (Table 1). The largest elbow angle found was 227°. In contrast, the vast majority of the κ light chains have elbow angles less than 180°, with a maximum elbow angle in this group of 196° (Table 1). Note that while a large percentage of λ chain Fab structures have elbow angles greater than 180°, λ chain Fabs can also display elbow angles as small as 117°, the smallest elbow angle found in this study. Thus, λ chain Fabs are not restricted to elbow angles >180°, but rather they are more able to assume larger (>195°) elbow angles compared to κ chain Fab fragments (Table 1; Supplementary Data, Table S1). A 2-way analysis of variance (ANOVA) of light chain type and heavy chain class reveals only the previously described significant correlation of elbow angle preference with light chain type, and no correlation with heavy chain class. There are insufficient data to allow a discrimination by species, as this distribution is dominated by mouse Fabs.

Fab elbow angles were examined to see if differences existed between the elbow angles of liganded and unliganded Fabs (Supplementary Data, Table S2). There are 61 Fabs in the dataset with structures for both their liganded and unliganded forms, many of these determined with multiple haptens, or in multiple crystal forms. Of these 61 Fabs, 38 show differences of less than 5° in elbow angle between the liganded and unliganded forms, with another 15 Fabs showing differences of between 5° and 20°. The largest elbow angle deviation was seen for the germline 48G7 (1AJ7 and 2RCS) with a 66° difference between the liganded and unliganded Fab. Since not all Fabs have different elbow angles for their free and bound forms, it seems likely that such changes are due to the inherent flexibility of the Fab, or to different crystal packing environments, as liganded and unliganded Fabs frequently crystallize in different crystal forms.

Amino acid sequence differences arise in the switch regions of λ *versus* κ light chains, including an inserted residue (L106a; Kabat numbering) in the sequence of λ light chains (Figure 3; Table 2). However, comparison of switch region structures from λ and κ light chains shows that residues

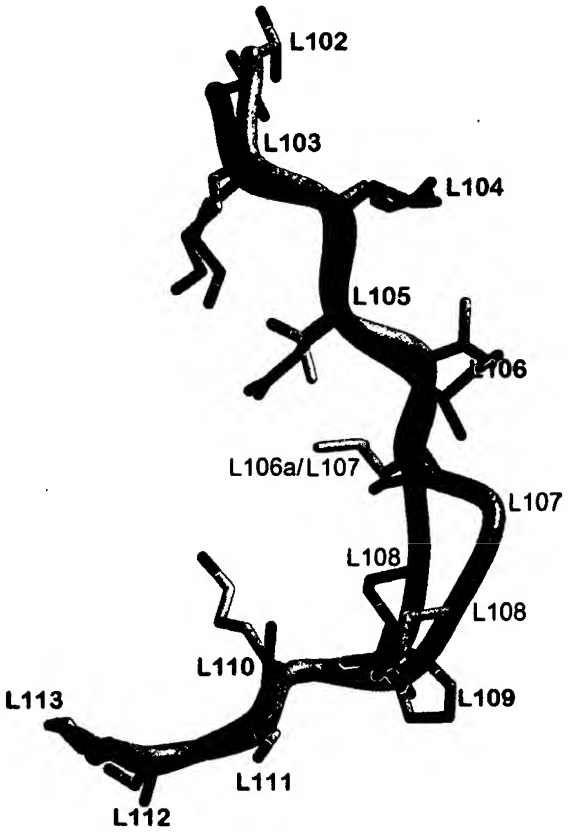


Figure 3. Switch regions from λ and κ light chains. The switch residues from λ (1ggc, light blue) and κ (1nj9, pink) light chains with similar elbow angles have been superimposed. The λ chains have an extra residue in the region (L106a by Kabat numbering); however, residue L107 (λ) is in fact the structural insertion, with residues L107(κ) and L106a(λ), and L108(κ) and L108(λ), being structurally equivalent.

L106a(λ) and L107(κ) are in fact structurally equivalent, as are L108(λ) and L108(κ), with residue L107(λ) corresponding to the insertion in the λ chain (Figure 3; Table 2), as defined by structural analysis. In λ chains, residue L107 is usually glycine (78.9% Gly, 13.4% Ser, 4.5% Arg from 960 λ sequences; 34 Gly, 1 Ser, 2 Arg in the 37 λ PDB entries). Switch regions are usually well-ordered in Fab crystal

Table 2. Amino acid preferences in elbow regions of κ and λ light chains

κ	λ
Leu/Val L104	Leu/Val L104
Glu/Asp L105	Thr L105
Ile/Leu/Val L106	Val L106
Lys L107	Leu L106a
	Gly L107
Arg L108	Gln L108
Ala/Thr L109	Pro L109

The λ chains bulge out at residue L107, so that κ L107 and λ L106a are structurally equivalent positions.

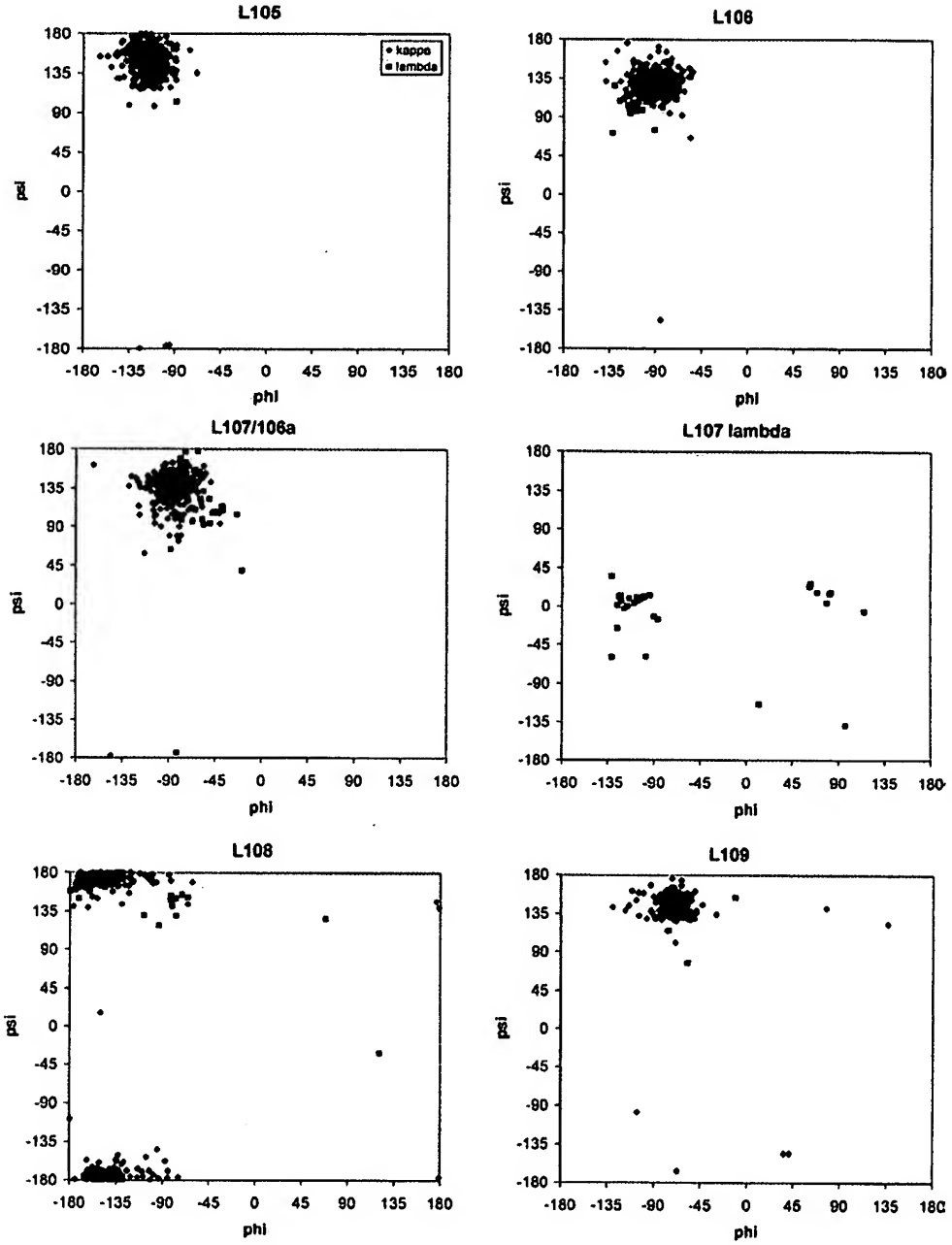


Figure 4. Ramachandran plots for residues in the light chain switch regions. Plots for residues L105 (a), L106 (b), λ L106a/κ L107 (c), λ L107 (d), L108 (e), and L109 (f). These plots combine both λ (pink) and κ (blue) chains, except for (d), where λ L107 is a structural insertion found only in λ chains.

structures, with only a few exceptions in the Fabs analyzed here, including the Fab with the smallest elbow angle (1jnh), which has no visible electron density for the switch region in any of the four Fabs in its asymmetric unit. The distributions of main-chain torsion angles for residues L105–L109 in the switch region are displayed in Figure 4 for all (λ and κ) Fab structures studied. Torsion angles are clustered fairly tightly for residues L105, L106, and L109, with more spread for L106a, L107, and L108, indicating that most of the elbow variation can be attributed to these residues. Comparison of κ and λ chain switch regions from Fabs with extreme elbow angle values (Figure 5) shows that large elbow angle differences in κ chains are manifested through consecutive small, sequential torsion angle changes around residues L105–L107, while λ chain Fabs exhibit more abrupt torsion angle changes centered around residues

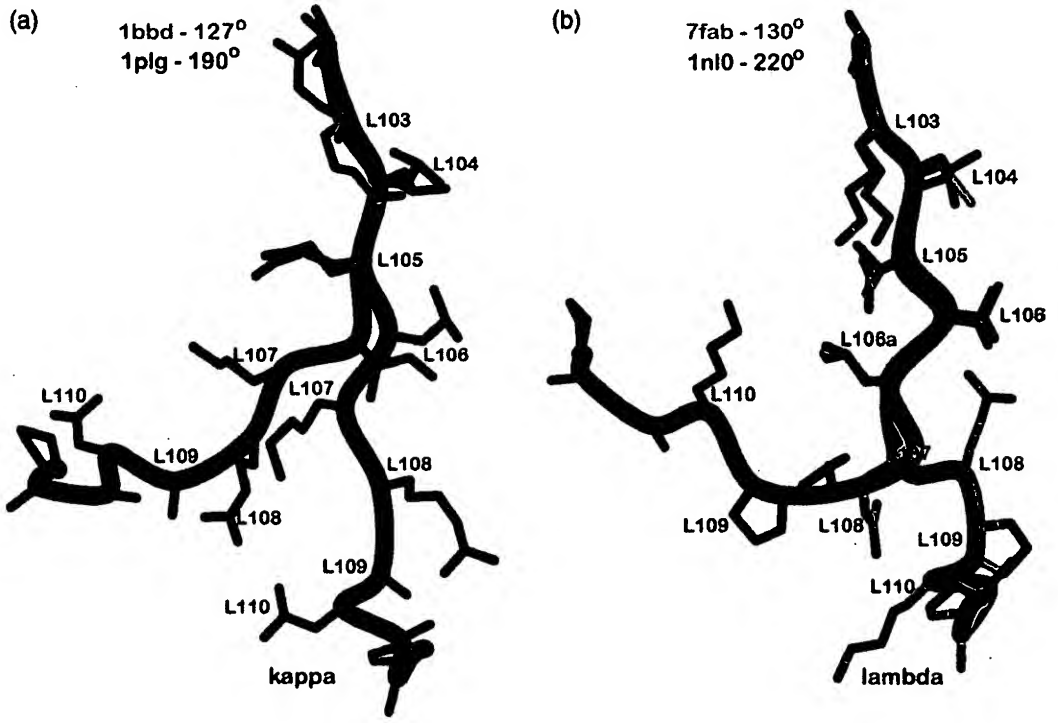


Figure 5. Comparison of extreme elbow angles from κ and λ chains. (a) The switch region residues from κ light chain Fabs 1bbd (127°, light blue) and 1plg (190°, blue) have been superimposed. This 63° difference is achieved by small movements around residues L105, L106, and L107. (b) Switch region residues from λ light chain Fabs 7fab (130°, light pink) and 1nl0 (220°, pink). This 90° difference is characterized by differences around residues L106a, L107 and L108, with the largest difference at residue L107. 1bbd and 1plg (mouse, IgG2a), and 7fab and 1nl0 (human, IgG1) have the same light and heavy chain constant region types.

L106a, L107 and L108. In this particular example, the largest difference between the two λ structures is at residue L107.

Early pioneering studies by Lesk & Chothia based on the available Fab crystal structures (only five at that time) had led to a proposal that Fab fragments could not achieve elbow angles greater than 180°.[12] In these five Fab crystal structures (New, McPC603, KOL, J539, and HyHEL5),[13–17] five conserved residues in the heavy chain were involved in key contacts between the V_H and C_H1 domains. These residues formed what was termed a flexible ball-and-socket joint, with Phe^{H148} and Pro^{H149} (Kabat numbering) serving as the ball that inserted into a socket formed by residues Leu/Val^{H11}, Thr^{H110}, and Ser^{H112} (Figure 6). No such conserved contact or interaction was found between the V_L and C_L domains of these structures. Despite different elbow angles in the Fabs, the five ball-and-socket residues always maintained physical contact, although they changed position slightly with respect to one another. Lesk & Chothia proposed that Fabs could not attain elbow angles of greater than 180° because the ball-and socket contacts would be lost in these extreme conformations.[12] However, in 1993, two crystal structures of an anti-chelate Fab (1ind, 1ine) unexpectedly revealed that Fab fragments could indeed adopt elbow angles greater than 180°, which in this case was ~194° for each of the two different ligand complexes.[18] Examination of the ball-and-socket region from structures of Fabs with vary large elbow angles shows that, as predicted, the ball-and-socket residues are not in contact (Figure 6(d)), although they remain fairly close in space. Thus, maintenance of the ball-and-socket contacts are not required, at least for Fabs with large elbow angles.

Conclusions

We have developed an easy to use, web-based service to calculate the elbow angle of a Fab fragment in a PDB format, and demonstrated its utility by rapidly calculating elbow angles for 536 Fab fragments in the PDB. The results show excellent agreement with previous compilations of Fab elbow angles.[6] The distribution of elbow angles is bimodal (Figure 2), with the largest elbow angles (>195°) found only in Fabs with λ light chains (the largest elbow angle seen for a κ light chain is 196°). These structural differences may be due to an additional residue (L107) in the λ chain switch region that perhaps allows for more flexibility. The L107 insertion is usually glycine, which also can provide more conformational freedom. It is not clear whether Fabs with λ light chains bend or flex around their elbow angle in solution more than

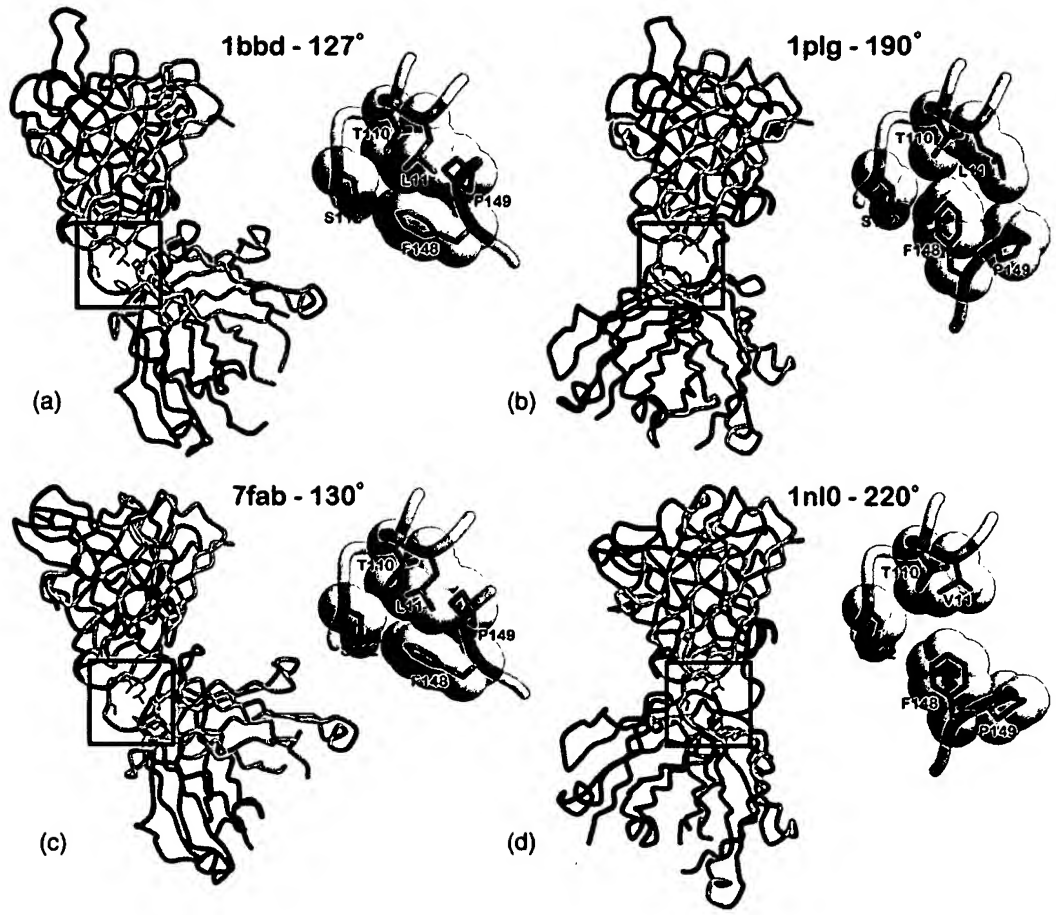


Figure 6. Ball-and-socket joint. The entire Fab fragment and a close-up of the residues that contribute to the heavy chain ball-and-socket joint are shown for Fabs with extreme elbow angles from both the κ and λ light chain class. (a) 1bbd; 127° elbow angle, murine IgG2a, κ (b) 1plg; 190° elbow angle, murine IgG2a, κ (c) 7fab; 130° elbow angle, human IgG1, λ (d) 1nl0; 220° elbow angle, human IgG1, λ.

other Fabs. Indeed, it is not certain how much elbow angle flexibility any particular Fab exhibits in solution. Molecular dynamics studies would indicate that elbow angle fluctuations of several degrees are common at least for κ chain Fabs in solution.[8] The availability of additional λ chain Fab structures in the future will likely allow for further refinement of our analysis. From a practical standpoint, the knowledge that λ chain Fabs tend towards large elbow angles may be useful to consider for the crystallographer carrying out molecular replacement structure determinations of these Fabs.

Acknowledgements

The authors gratefully acknowledge Dr Marc Elsliger's help in establishing the web-server, and NIH grants GM-46192 and GM-38273 for support to R.L.S and I.A.W. Part of the work was performed under the auspices of the US Department of Energy by the University of California, Lawrence Livermore National Laboratory under contract W-7405-Eng-48. This is manuscript number 17618-MB from the Scripps Research Institute.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2006.01.023

References

1. Schiffer, M., Girling, R. L., Ely, K. R. & Edmundson, A. B. (1973). Structure of a lambda-type Bence-Jones protein at 3.5-Å resolution. *Biochemistry*, **12**, 4620–4631.
2. Huber, R., Deisenhofer, J., Colman, P. M., Matsushima, M. & Palm, W. (1976). Crystallographic structure studies of an IgG molecule and an Fc fragment. *Nature*, **264**, 415–420.
3. Kabat, E. A., Wu, T. T., Perry, H. M., Gottesman, K. S. & Foeller, C. (1991). *Sequences of Proteins of Immuno-*

logical Interest. 5th edit., U.S. Department of Health and Human Services.
4. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al.* (2000). The Protein Data Bank. *Nucl. Acids Res.* **28**, 235–242.
5. Zemla, A. (2003). LGA: a method for finding 3D similarities in protein structures. *Nucl. Acids Res.* **31**, 3370–3374.
6. Wilson, I. A. & Stanfield, R. L. (1994). Antibody-antigen interactions: new structures and new conformational changes. *Curr. Opin. Struct. Biol.* **4**, 857–867.
7. Rossmann, M. G. & Argos, P. (1975). A comparison of the heme binding pocket in globins and cytochrome *b5*. *J. Biol. Chem.* **250**, 7525–7532.
8. Sotriffer, C. A., Rode, B. M., Varga, J. M. & Liedl, K. R. (2000). Elbow flexibility and ligand-induced domain rearrangements in antibody Fab NC6.8: large effects of a small hapten. *Biophys. J.* **79**, 614–628.
9. Kantardjieff, K. A. & Rupp, B. (2003). Matthews coefficient probabilities: improved estimates for unit cell contents of proteins, DNA, and protein-nucleic acid complex crystals. *Protein Sci.* **12**, 1865–1871.
10. Wukovitz, S. W. & Yeates, T. O. (1995). Why protein crystals favour some space-groups over others. *Nature Struct. Biol.* **2**, 1062–1067.
11. Almagro, J. C., Hernandez, I., Ramirez, M. C. & Vargas-Madrazo, E. (1998). Structural differences between the repertoires of mouse and human germ-line genes and their evolutionary implications. *Immunogenetics*, **47**, 355–363.
12. Lesk, A. M. & Chothia, C. (1988). Elbow motion in the immunoglobulins involves a molecular ball-and-socket joint. *Nature*, **335**, 188–190.
13. Saul, F. A., Amzel, L. M. & Poljak, R. J. (1978). Preliminary refinement and structural analysis of the Fab fragment from human immunoglobulin New at 2.0 Å resolution. *J. Biol. Chem.* **253**, 585–597.
14. Satow, Y., Cohen, G. H., Padlan, E. A. & Davies, D. R. (1986). Phosphocholine binding immunoglobulin Fab McPC603. An X-ray diffraction study at 2.7 Å. *J. Mol. Biol.* **190**, 593–604.
15. Marquart, M., Deisenhofer, J., Huber, R. & Palm, W. (1980). Crystallographic refinement and atomic models of the intact immunoglobulin molecule Kol and its antigen-binding fragment at 3.0 Å and 1.9 Å resolution. *J. Mol. Biol.* **141**, 369–391.
16. Suh, S. W., Bhat, T. N., Navia, M. A., Cohen, G. H., Rao, D. N., Rudikoff, S. & Davies, D. R. (1986). The galactan-binding immunoglobulin Fab J539: an X-ray diffraction study at 2.6-Å resolution. *Proteins: Struct. Funct. Genet.* **1**, 74–80.
17. Sheriff, S., Silverton, E. W., Padlan, E. A., Cohen, G. H., Smith-Gill, S. J., Finzel, B. C. & Davies, D. R. (1987). Three-dimensional structure of an antibody-antigen complex. *Proc. Natl Acad. Sci. USA*, **84**, 8075–8079.
18. Love, R. A., Villafranca, J. E., Aust, R. M., Nakamura, K. K., Jue, R. A., Major, J. G., Jr *et al.* (1993). How the anti-(metal chelate) antibody CHA255 is specific for the metal ion of its antigen: X-ray structures for two Fab′/hapten complexes with different metals in the chelate. *Biochemistry*, **32**, 10950–10959.

Edited by Michael J. E. Sternberg

(Received 21 November 2005; received in revised form 1 January 2006; accepted 4 January 2006)
Available online 25 January 2006

electronic reprint

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449
Editors: **E. N. Baker and Z. Dauter**

Mycobacterium tuberculosis RmlC epimerase (Rv3465): a promising drug-target structure in the rhamnose pathway

Katherine A. Kantardjieff, Chang-Yub Kim, Cleo Naranjo, Geoffry S. Waldo, Timothy Lekin, Brent W. Segelke, Adam Zemla, Min S. Park, Thomas C. Terwilliger and Bernhard Rupp

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Mycobacterium tuberculosis RmlC epimerase (Rv3465): a promising drug-target structure in the rhamnose pathway

Katherine A. Kantardjieff,[a] Chang-Yub Kim,[b] Cleo Naranjo,[b] Geoffry S. Waldo,[b] Timothy Lekin,[c] Brent W. Segelke,[c] Adam Zemla,[c] Min S. Park,[b] Thomas C. Terwilliger[b] and Bernhard Rupp[c]*

[a]Department of Chemistry and Biochemistry and W. M. Keck Foundation Center for Molecular Structure, California State University Fullerton, Fullerton, CA 92834, USA, [b]Bioscience Division, MS M888, Los Alamos National Laboratory, Los Alamos, NM 87545, USA, and [c]Biology and Biotechnology Research Program, L-448, University of California, Lawrence Livermore National Laboratory, Livermore, CA 94551, USA

Correspondence e-mail: br@llnl.gov

Received 17 February 2004
Accepted 5 March 2004

PDB Reference: dTDP-4-keto-6-deoxyglucose epimerase, 1upi, r1upisf.

The *Mycobacterium tuberculosis rmlC* gene encodes dTDP-4-keto-6-deoxyglucose epimerase, the third enzyme in the *M. tuberculosis* dTDP-L-rhamnose pathway which is essential for mycobacterial cell-wall synthesis. Because it is structurally unique, highly substrate-specific and does not require a cofactor, RmlC is considered to be the most promising drug target in the pathway, and the *M. tuberculosis rmlC* gene was selected in the initial round of TB Structural Genomics Consortium targets for structure determination. The 1.7 Å native structure determined by the consortium facilities is reported and implications for *in silico* screening of ligands for structure-guided drug design are discussed.

1. Introduction

The TB Structural Genomics Consortium (TBSGC) is one of the nine NIGMS-funded Protein Structure Initiative Pilot projects serving as a structural biology resource for the *Mycobacterium tuberculosis* (MTB) research community (Terwilliger *et al.*, 2003). Consortium members can target proteins of interest, and highly ranked targets are produced at the Los Alamos protein-production facility and shipped to the crystallization facility at Lawrence Livermore National Laboratory (LLNL) for automated high-throughput crystallization (Rupp *et al.*, 2002). Data are collected at the Advanced Light Source (ALS) in Berkeley, and the structures are determined by the LLNL or ALS beamline members. Coordinates are deposited within three weeks of the final refinement.

TB is a re-emerging disease with an increasing prevalence of multi-drug resistant strains (Ramaswamy & Musser, 1998), and a long-term goal of the TBSGC is to provide a foundation for structure-guided drug design. Protein targets of high priority are those which are essential or unique to the bacillus. Mycobacterial cell-wall biosynthesis, the target of the well known drugs isoniazid and ethambutol (Schroeder *et al.*, 2002), has been of particular interest in the development of antimycobacterial therapeutics (Kantardjieff & Rupp, 2004). Rhamnose-synthetic enzymes are attractive targets in cell-wall synthesis. L-Rhamnose, a sugar that is not present in the human host, plays a key role as a structural link between the mycobacterial cell-wall components arabinogalactan and peptidoglycan. L-Rhamnose is derived from a precursor, dTDP-L-rhamnose, which is synthesized from glucose-1-phosphate and dTTP by means of the dTDP-L-rhamnose pathway. The prokaryotic dTDP-L-rhamnose pathway consists of four enzymes, the three-dimensional structures of each of which have been determined from different bacteria.

RmlC, the third enzyme in the dTDP-L-rhamnose pathway, functions as an epimerase, converting dTDP-4-keto-6-deoxy-

Table 1
Data-collection and refinement statistics.

Values in parentheses are for the highest resolution bin (1.75–1.70 Å).

Data collection	
Space group	$P3_221$
Wavelength (Å)	1.000
Unit-cell parameters	
a, b (Å)	64.91
c (Å)	87.20
Resolution (Å)	55.90–1.7
Unique reflections	23837 (1839)
Redundancy	4.4 (4.0)
Completeness	99.2 (93.8)
R_{sym}	0.052 (0.410)
$(I/\sigma(I))$	46.8 (2.5)
Reflections with $I/\sigma(I) > 3$ (%)	87.8 (53.3)
No. molecules in AU	1
V_M (Å^3 Da^{-1})	2.14
Solvent content (%)	42.7
Refinement	
R_{free} value, random 5% of reflections	0.249 (0.340)
R value	0.201 (0.272)
R.m.s.d. bond lengths† (Å)	0.021
R.m.s.d. bond angles† (°)	1.786
Overall coordinate error‡ (Å)	0.127
RSCC (*Shake&wARP*)§	0.92
RSCC (*REFMAC5*)¶	0.96
Ramachandran appearance††, residues in	
Most favored region	155 (89.6%)
Additional allowed	17 (9.8%)
Generously allowed	1 (0.6%)
Disallowed	0

† Deviations from restraint targets (Engh & Huber, 1991). ‡ Estimated standard uncertainty: diffraction precision index (DPI) based on R_{free} (Cruickshank, 1999). § Real-space correlation coefficient, F_c map against averaged and weighted *Shake&wARP* map. ¶ Real-space correlation coefficient, F_o map against F_c map, as reported by *REFMAC5*. †† Regions as defined in *PROCHECK* (Laskowski *et al.*, 1993).

glucose to dTDP-4-keto-rhamnose (Stern *et al.*, 1999). Because it is structurally unique, highly substrate-specific and does not require a cofactor, RmlC is considered to be the most promising drug target in the pathway, and the MTB *rmlC* gene was selected in the initial round of consortium targets for structure determination. The structures of RmlC from *Methanobacterium thermoautotrophicum*, *Salmonella typhimurium* and *Streptococcus suis*, both uncomplexed and bound to substrate analogs, have also been determined recently (Christendat *et al.*, 2000; Giraud *et al.*, 2000; Babaoglu *et al.*, 2003). We present a brief structure description, a comparison with the other uncomplexed apo structures and results from virtual ligand-screening studies of known and potential inhibitors. The coordinates (PDB code 1upi) have been deposited and released immediately in accordance with NIH guidelines for Structural Genomics Pilot Projects.

2. Experimental methods

2.1. Cloning and expression

A 0.6 kbp DNA fragment containing the *rmlC* gene (EMBL locus MTY13E12, accession No. Z95390.1, Rv3465) was amplified from *Mycobacterium tuberculosis* H37Rv genomic DNA as the PCR template, using the following oligonucleotide primers: 5′-AGATATA**CATATG**AAAGCACGCGAACTCGACGTCCCC-3′ and 5′-AATTC**GGATCC**GGTGCCGCGCATCTCCCCAATGAA-3′. The bases in bold represent *Nde*I and *Bam*HI sites, respectively. The amplified DNA fragment was digested with *Nde*I and *Bam*HI restriction enzymes and subcloned into the corresponding restriction sites in a modified pET28b vector which provided an N-terminal six-His tag upstream of the *Nde*I site. The expressed protein, thus has the N-terminal extension MGSSHHHHHHSSGLVPRGSH and an additional GSV at the C-terminus.

Escherichia coli BL21 (DE3) cells were transformed with the *rmlC*-modified pET28b/His vector and grown to exponential phase at 310 K in 5 ml EZMix LB broth medium (Sigma) containing 30 µg ml^{-1} kanamycin and 50 µg ml^{-1} spectinomycin. This seed culture was transferred to 0.5 l EZMix Terrific broth medium (Sigma) and expression was induced with 0.5 m*M* IPTG at an OD_{600} of approximately 0.5. Growth was continued at 293 K for approximately 21 h until the OD_{600} reached approximately 15 (as inferred from dilutions). The cells were harvested and stored at 193 K.

2.2. RmlC purification

The cell pellet was lysed by sonication in 10 ml buffer *A* (20 m*M* Tris pH 8.0, 100 m*M* NaCl) per gram of cells for 10 min in 30 s pulses at 283 K. The cell debris was removed by ultracentrifugation for 30 min at 38 000 rev min^{-1} using a Ti-60 rotor (Beckman). The clear supernatant was filtered through a 0.2 µm pore membrane and loaded onto a 5 ml Talon superflow affinity column equilibrated with buffer *A*. After washing with 50 ml buffer *A*, the His-tagged RmlC was eluted from the cobalt-affinity column using buffer *B* (20 m*M* Tris–HCl pH 8.0, 500 m*M* NaCl and 300 m*M* imidazole). The elutant was dialyzed against buffer *C* (20 m*M* Tris–HCl pH 8.0, 100 m*M* NaCl, 10 m*M* β-mercaptoethanol) and purified by gel filtration on a Superdex-75 column using buffer *C* for equilibration and elution (Amersham Pharmacia Biotech). The peak fractions (monitored at OD_{280}) were analyzed by SDS–PAGE and the pooled protein fractions were concentrated to 26 mg ml^{-1} using a Centriprep YM-3 (Millipore). Protein, the purity of which was estimated to be 99% by SDS–PAGE and MALDI–TOF mass spectroscopy (Applied Biosystems), was stored at 277 K and shipped to the TB consortium crystallization facility at LLNL (Rupp *et al.*, 2002).

2.3. Crystallization

Crystals were grown in Greiner 96-well plates from sitting drops consisting of 2 µl protein stock solution mixed with 2 µl well solution. Conditions were screened using the CRYSTOOL random-screening protocol (Segelke, 2001) and the first crystals were observed one week after setup. Of the 288 precipitant conditions tested, 0.1 *M* sodium citrate buffer pH 5.5, 28% PEG MME 2K and 0.33% LDAO yielded diffraction-quality crystals.

2.4. Data collection

A rhomboid crystal of approximately 50 µm in size in all dimensions was harvested in a Hampton cryoloop and immediately flash-cooled in liquid nitrogen. The cryopin was placed in a puck of the Advanced Light Source (ALS) storage and transfer system (Rupp *et al.*, 2002) and robotically mounted on ALS beamline 5.0.3. Data to 1.65 Å were collected at an X-ray wavelength of 1.000 Å, integrated using *HKL*2000, and scaled with *SCALEPACK* (Otwinowski & Minor, 1997) in trigonal Laue group $\bar{3}$. The data were further reduced in final space group $P3_221$ (No. 154), with unit-cell parameters a = 64.91, c = 87.20 Å. Calculation of Matthews probabilities (Matthews, 1968; Kantardjieff & Rupp, 2003) and solvent density indicated there to be one molecule in the asymmetric unit. Data-collection statistics are summarized in Table 1 and details are available from the PDB header. Successful molecular replacement established $P3_221$ as the correct selection from the enantiomorphic pair (No. 154 *versus* No. 152).

2.5. Structure determination

The structure of MTB RmlC was determined by molecular replacement using a homology model built with the automated protein-structure prediction system AS2TS (Amino-acid Sequence to Tertiary Structure) developed at LLNL (Rupp *et al.*, 2002). The dTDP-4-dehydrothamnose 3,5-epimerase RmlC from the archaeon *Methanobacterium thermoautotrophicum* (PDB code 1ep0, chain *A*; 183 residues; Christendat *et al.*, 2000), which shares an identity of 38% over 185 residues with MTB RmlC (Fig. 1), was used as a template to calculate

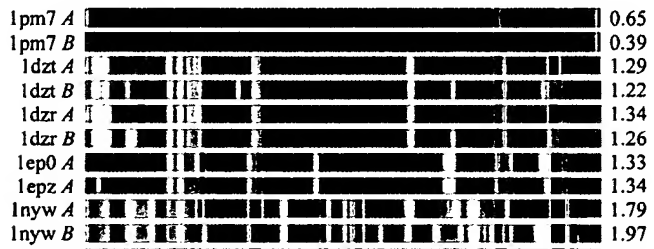


Figure 1
Pairwise structural alignment of homologous protein chains with RmlC from MTB using the local global alignment program *LGA* (Zemla, 2003). The left column indicates the PDB code and chain ID and colored bars represent $C^{\alpha}-C^{\alpha}$ distance deviation between superimposed PDB structures and RmlC (200 residues; from the N-terminus at the left to the C-terminus at the right). Residues superimposed below 1.5 Å are in green, below 3.0 Å in yellow, below 4.5 Å in orange, below 6.0 Å in brown and residues at or above 6.0 Å are in red. Non-aligned terminal residues are in gray. The right column contains r.m.s.d.s in Å calculated for all C^{α} pairs that are superimposed under 5 Å distance cutoff. The plot shows that homologous proteins differ significantly (red) from TB RmlC in the C-terminal part (loop 160–165, region 179–186) and also that the C-terminal helix is not present (gray) in the templates.

the main-chain atoms in the model; side-chain atoms were calculated using the program *SCWRL* (Canutescu *et al.*, 2003). *EPMR* (Kissinger *et al.*, 1999) was used with default settings (15–4 Å, no bump restraints), and searches converged at correlation coefficients of 0.33, with an R value of 0.48 after rigid-body refinement against data to 2.8 Å.

2.6. Model building and refinement

To ensure effective phase-bias removal, the model was iteratively built using the program *XFIT* (McRee, 1999) into maps generated by the *Shake&wARP* procedure implemented in the TB consortium map-improvement server (Reddy *et al.*, 2003). An additional C-terminal helix was clearly visible in the initial map, as were several major loop arrangements and residue modifications. After repeated cycles of solvent building and real-space refinement, followed by restrained *REFMAC5* maximum-likelihood refinement (Murshudov *et al.*, 1997), the final structure (PDB code 1upi) refined to R = 0.201 and $R_{\rm free}$ = 0.249. Weak density for three additional residues from the N-terminal His tag was visible in the maps, but these could not be reliably modeled and have been omitted from the model. At the C-terminus, the last two residues from the protein and three additional residues from a *Bam*H1 cloning artefact were also absent. Real-space correlation coefficient plots ($\langle CC \rangle$ = 0.92) have been calculated by the TB consortium map-improvement server (Reddy *et al.*, 2003). Details of the refinement and data-collection statistics are tabulated in the header file of PDB entry 1upi and are briefly summarized in Table 1.

2.7. Quality assessment

PROCHECK (Laskowski *et al.*, 1993) and *WHAT_CHECK* (Hooft *et al.*, 1996) reports were created upon coordinate deposition and are available from the PDB for entry 1upi. Ramachandran plot distribution, coordinate error, r.m.s.d. from target-geometry values and real-space correlation which are typical for a well refined 1.7 Å structure are summarized in Table 1. The only residue in a generously allowed region of the Ramachandran plot region is Ala161, which is located in a disordered loop.

2.8. *In silico* virtual ligand screening

To virtually screen for potential inhibitors of MTB RmlC, flexible docking simulations were performed with *ICM-Pro* 3.0.25l (Schapira *et al.*, 2003). To test the robustness of the docking procedure, crystal structures of RmlC ligand complexes from *Streptococcus suis*, *Salmonella typhimurium* and *Methanobacterium thermoautotrophicum* were simulated first. Similar substrate analogs as well as reported active compounds (Andres *et al.*, 2000; Ma *et al.*, 2001; Babaoglu *et al.*, 2003) were then docked to MTB RmlC using superposition with a ligand-bound structure and inferences from sequence alignment to initially define the receptor site. The active sites in the MTB RmlC dimer were also characterized using ICMPocketFinder, which detects cavities of sufficient size to bind 'druggable' molecules.

3. Results

3.1. Structure summary

MTB RmlC is an obligate homodimer in which the dimer axis in 1upi coincides with the crystallographic twofold. The buried interface excludes 1500 Å^2 per molecule (Fig. 2). The overall topology of MTB RmlC is consistent with the structural classification of these proteins, mainly β-class with a jelly-roll-like topology, in which each monomer is characterized by a double-stranded β-helix forming the active site of the enzyme (Laskowski, 2001). As seen in the other dimer structures, β-strands extending from each monomer stabilize the dimer. The extended strand from the N-terminus of one monomer contributes residues to the active site of the other (Christendat *et al.*, 2000; Giraud *et al.*, 2000; Babaoglu *et al.*, 2003). The most significant difference between MTB RmlC and the other RmlC structures is a well defined C-terminal ten-residue helix extension. The surface-accessible cysteine residues Cys134 and Cys146 could be clearly modeled as *S*,*S*-(2-hydroxyethyl)-thiocysteine (CME; Fig. 3), presumably resulting from modification by β-mercaptoethanol present in the dialysis buffer. The buried cysteine residues Cys50 and Cys76 are not modified.

The structure contains two non-proline *cis*-peptides, Gly60–Leu61, which is clearly defined (Fig. 4), and Asp161–Gly162, which is located in a disordered loop region. *Cis*-peptide Gly60–Leu61, part of the highly conserved sequence VLRGLH, has been observed in all reported RmlC structures and is likely to have biological relevance (Weiss & Hilgenfeld, 1999), as it forms part of the active-site binding pocket. It has been proposed that because this energetically unfavorable conformation is highly conserved, Gly60 may help to orient catalytic residues on β6 in the active site (Christendat *et al.*, 2000). An exception to this *cis* conformation is found in 1dzr (*Salmonella typhimurium*), in which the same residues display a large deviation from geometry targets, indicating uncertainty

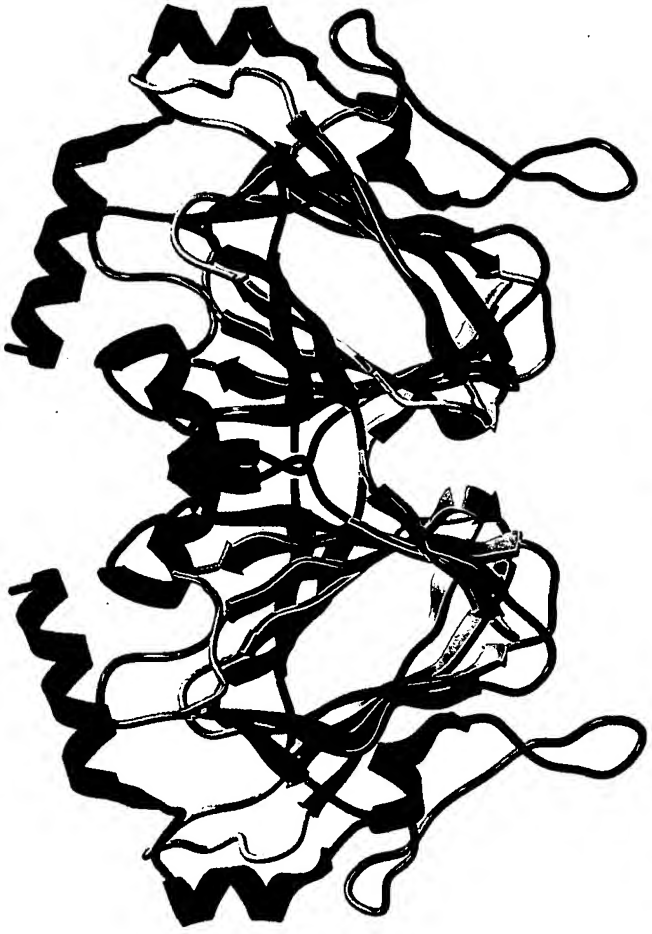


Figure 2
MTB RmlC homodimer. Ribbon drawing showing the homodimer and jelly roll-like topology, in which each monomer (colored by residue number and varying from blue at the N-terminus to red at the C-terminus) is characterized by a double-stranded β-helix forming the active site of the enzyme (Laskowski, 2001). An extended β-strand (shown in yellow) from the N-terminus of each monomer contributes residues to the active site of the other (Christendat *et al.*, 2000; Giraud *et al.*, 2000; Babaoglu *et al.*, 2003). The view looks into the opening of active site of the lower monomer. The image was rendered with *ICM-Pro* v.3.0.25l.

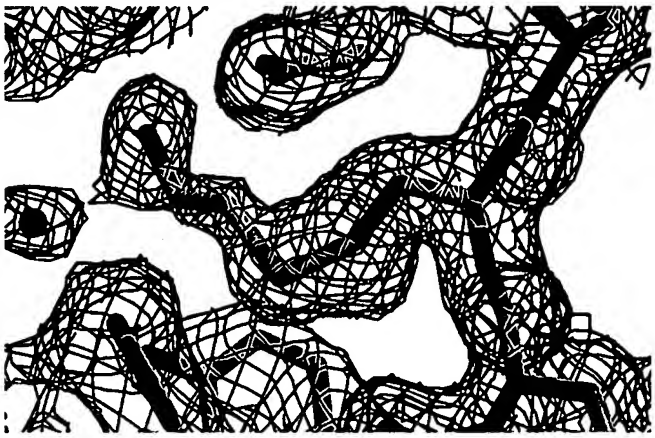


Figure 3
Modified solvent-exposed *S*,*S*-(2-hydroxyethyl)-thiocysteine (CME). The electron-density map around CME134 was generated by the TB consortium bias-removal and map-improvement server (Reddy *et al.*, 2003) and is contoured at the 0.7σ level. The figure was created with *XFIT* (McRee, 1999) and rendered with *RASTER3D* (Merritt & Bacon, 1997).

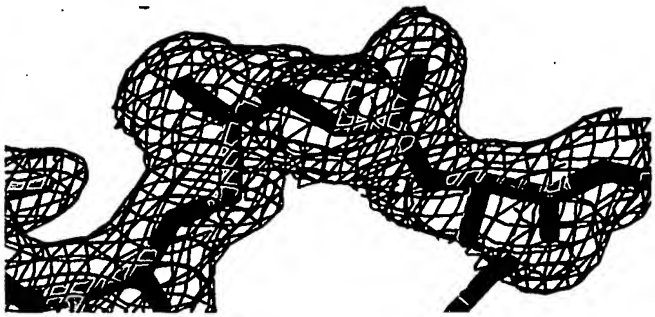


Figure 4
The structurally and sequentially conserved *cis*-peptide Gly60–Leu61 in MTB RmlC (PDB code 1upi). Electron density was calculated by the TB consortium map-improvement server (Reddy *et al.*, 2003) and is contoured at 1σ. The side chains of Leu61 and His62 are clipped for better backbone clarity. The figure was created with *XFIT* (McRee, 1999) and *RASTER3D* (Merritt & Bacon, 1997).

Table 2
Active-site binding-pocket characteristics for MTB RmlC.

PDB code	Volume (Å^3)	Area (Å^2)	Chain *B* residues	Chain *A* residues
1upi chain *A*	541	486	19, 23, 26, 28	47, 49, 51, 59, 62, 70, 72, 119–121, 132, 134, 138, 140, 143–145, 168, 170–171, 174–175
1dzr chain *A*	588	757	2, 15, 17–18, 20–21, 24	27, 29–32, 35–36, 74, 111; 48, 52, 54–55, 58, 60–61, 63, 65–66, 73, 120–122, 131, 133, 139, 144–146, 167, 169–170, 174, 178–179, 181–182
1ep0 chain *A*	535	456	3, 22–24, 26, 28–29, 31	51, 53, 61–62, 64–66, 71, 73, 120–122, 133, 139, 144–145, 169, 171–172, 175
1nxm chain *A*	392	420	29–31, 33–36, 38	61, 63, 65–68, 71, 73–74, 76, 78, 80, 82, 127–129, 138, 140, 142, 146, 175

in the modeling. Inspection of the bias-minimized electron density obtained from the TB consortium bias-removal server (Reddy *et al.*, 2003) suggests that Gly60–Leu61 in 1dzr could in fact be modeled in a *cis* conformation. The *cis*-peptide detection program *ANGL_ANAL* (Weiss & Hilgenfeld, 1999) also ranks the Gly60–Leu61 peptide bond in 1dzr as *cis* with a high probability score.

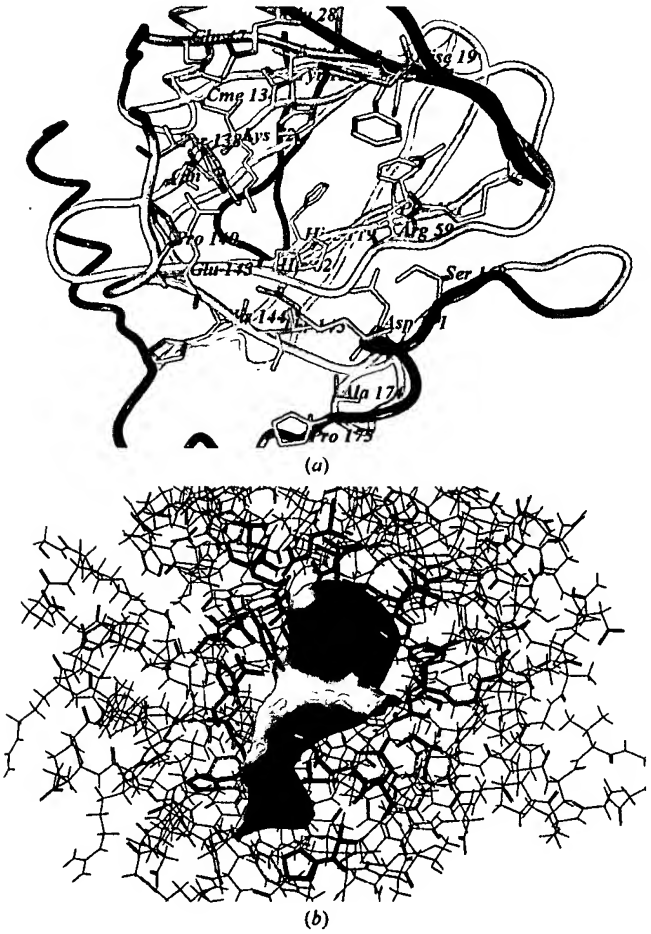


Figure 5
Active site of MTB RmlC. (*a*) Binding-pocket residues as indicated by ICMPocketFinder and reported in Table 2. The polypeptide chain is colored from blue at the N-terminus to red at the C-terminus for each monomer. Yellow sticks represent protein residues describing the pocket. The view is the same as in Fig. 2. (*b*) Binding-pocket volume as defined by ICMPocketFinder for virtual ligand screening. The gray mesh denotes the ligand-accessible pocket volume (500 Å^3) of the protein receptor. Yellow sticks represent protein residues describing the pocket. The view is the same as in (*a*).

Loop Leu159–Ala163 is also disordered in a 2.2 Å MTB RmlC dimer structure (PDB code 1pm7, released during revision of this manuscript) and presents the only major difference between the two TB structures. The total backbone C$^\alpha$-atom r.m.s.d. between each of the two monomers in 1pm7 and the single monomer unit in 1upi excluding Met1, Leu159–Ala163 and the C-terminal Gly197–Met199 are 0.30 and 0.26 Å, respectively. Omitting the same residues in a superposition of the two chains of 1pm7 onto each other yields an r.m.s.d. of 0.16 Å, less than the expected Cruickshank coordinate error (Cruickshank, 1999) of 0.23 Å (1pm7). Together with the very small deviations of the superposition (NCS) matrix from a pure twofold operator perpendicular to axis *c* (0.04°, 0.16 Å displacement along *c*), we conclude that the two structures are highly related. It is also consistent that the higher symmetry structure (1upi) diffracts to higher resolution (1.7 *versus* 2.2 Å in 1pm7). The possibility of significantly different dimer orientations in the two structures can also be excluded, as both dimers superimpose with a total C$^\alpha$ r.m.s.d. of 0.35 Å. Different protein constructs, purification buffers and crystallization conditions are possible sources of the differences in crystal form.

Details of the MTB RmlC active site are shown in Fig. 5. Highly conserved within the MTB RmlC active site are the His–Asp dyads His119–Asp83 and His62–Asp171 as well as Lys72, all of which are ionizable groups strategically placed to participate in acid/base chemistry within the active site (Christendat *et al.*, 2000). Phe121 and Tyr138, which have been implicated in carbohydrate binding (Babaoglu *et al.*, 2003), and the hydrophilic residues Gln47, Asn49, Ser1 and Ser53, which comprise a network for substrate binding and catalysis (Christendat *et al.*, 2000), are also highly conserved. Table 2 reports the characteristics of the active sites in available RmlC dimers, as determined by ICMPocketFinder. The greatest differences are seen in 1dzr (*Salmonella typhimurium*), the total backbone C$^\alpha$-atom r.m.s.d. of which deviates significantly from other RmlCs and the active-site binding pocket of which is distinctly smaller in volume than the other apo RmlC structures. The binding pocket of MTB RmlC is intermediate in volume and surface area between 1dzr (*Salmonella typhimurium*, smallest) and 1nxm (*Streptococcus suis*, largest) and is most similar to 1ep0 (*Methanobacterium thermoauto-*

trophicum). Yet, despite the observed differences in structural details, the sequence and structural elements essential to function are highly conserved among RmlCs, particularly within the double-stranded β-helix forming the active site of the enzyme (Fig. 6).

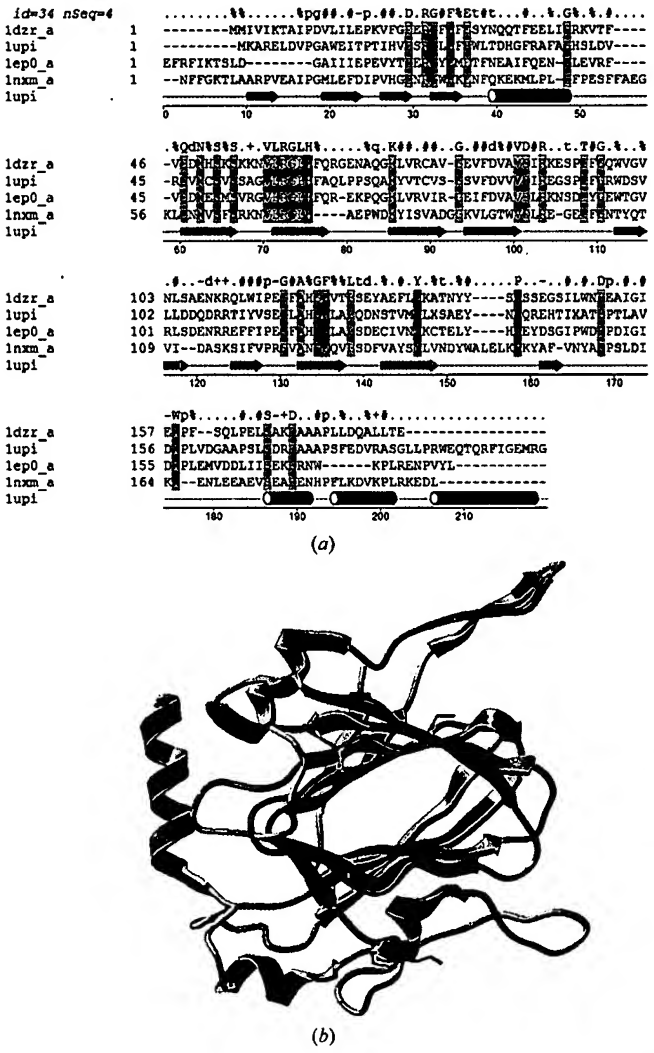


Figure 6
Combination sequence and structural alignment colored by consensus strength (strong in dark green to weak in white). (*a*) Sequence alignment of available RmlC monomer structures. The secondary-structure scheme for MTB RmlC is shown below the alignment and the consensus sequence is shown above the alignment. The consensus string contains the following symbols: +, positively charged amino acids R and K; −, negatively charged amino acids D and E; (^) small amino acids A, S and G; %, aromatic residues F, Y and W; #, hydrophobic amino acids F, I, L, M, P, V and W; ~, polar amino acids C, D, E, G, H, N, Q, S, T, Y; dot, no consensus, no gap. For example, if Gly is found in more than 85% of sequences its consensus symbol is 'G'; if the percentage is between 60–85 the symbol becomes 'g'; if no consensus is established, the symbol becomes '.'. For residues WLVIMAFCYHP, '#' indicates that the residue is found in more than 85% of the sequences and '%' if the percentage is between 60 and 85%. (*b*) Ribbon diagram of MTB RmlC monomer, colored by consensus strength as in (*a*). View is from the bottom of the monomer, the same view as in Fig. 2. Calculated and rendered with *ICM-Pro* (Abagyan *et al.*, 1994, 1997).

3.2. Virtual screening

The ligand conformations in crystal structures of the dimeric RmlC–ligand complexes (*Streptococcus suis*, *Salmonella typhimurium* and *Methanobacterium thermoautotrophicum*) could be reproduced *in silico* within 2–3 Å r.m.s.d., which is considered to be the benchmark for successful simulated docking (Totrov & Abagyan, 2001). Subsequent simulated docking of the same ligands to MTB RmlC produced conformations that closely resembled those observed in the crystal structures of homologs. These simulated complexes were consistently among the top-scoring conformations and were distinguishable from less specific binding modes in that the pyrimidine rings tended to be superimposable; the substrate-binding mode exhibited by these enzymes is a ring π-stacking (Christendat *et al.*, 2000; Giraud *et al.*, 2000; Babaoglu *et al.*, 2003) as seen in the nucleotide-binding regions of many other proteins. Furthermore, although the sugar moiety moves within the active site, as has been observed in crystal structures (Babaoglu *et al.*, 2003), the diphosphates locate themselves similarly, interacting with several conserved arginines. The simulated docking results suggest that the relative superposition of putative inhibitors/analog compounds with pyrimidine and phosphate functionalities should be important factors for drug design. Indeed, ICM successfully docked reported active classes of compounds (Ma *et al.*, 2001; Babaoglu *et al.*, 2003),

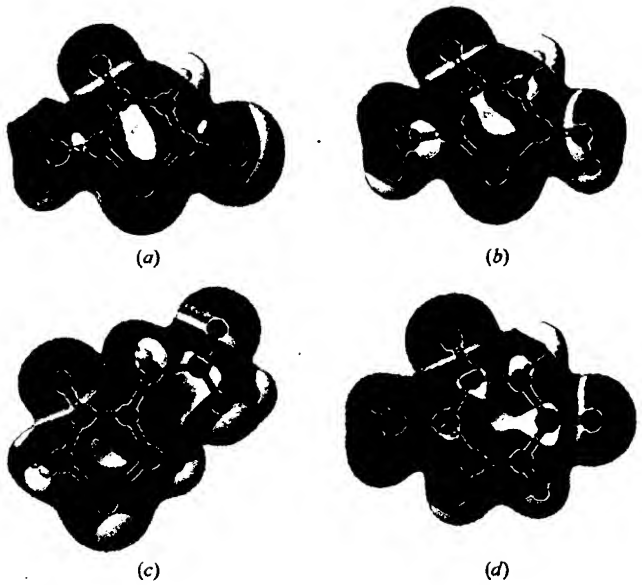


Figure 7
Core structure electron-density isosurfaces for antimycobacterial compounds and substrate analogs colored by electrostatic potential. The potential at a point near a molecule is the potential energy of a positive charge at that point. Coloring is from white (highest, positive) to purple (lowest, negative) calculated with *CaChe WorkGroup Pro* v.6.1.1 at the B88-LYP DFT level (Becke, 1988; Lee *et al.*, 1988). (*a*) Rhodanine core; (*b*) rhodanine-like core; (*c*) thiazolidinone; (*d*) thymine. Compounds with rhodanine or thiazolidinone core structures (*a* and *c*) tended to dock at or near the phosphate-binding region of MTB RmlC, whereas compounds with rhodanine-like core structures (*b*) docked in or near the nucleotide-binding pocket (*d*).

including several compounds from a limited Nanosyn library meeting Lipinski's criteria (Lipinski *et al.*, 2001), with surrogate pharmacophores correspondingly disposed in high-scoring orientations.

As a follow-up to the work of Ma *et al.* (2001), we docked to MTB RmlC a series of compounds containing a rhodanine (van der Helm *et al.*, 1962) or rhodanine-like core structure, which had shown activity against MTB RmlC and inhibited growth of MTB in culture and which the authors had suggested to provide preliminary structure–activity relationships. Consistently, the rhodanine-like core structure bound in or near the nucleotide-binding region by analogy with homologous structures, whereas rhodanine and thiazolidinone core structures bound at or near the phosphate-binding region. The binding preferences for these structures may be explained by their electrostatic potential surfaces (Fig. 7). Substituted thiazolidinones share common features with compounds thought to mimic the diphosphate moiety in the transition state (Biller *et al.*, 1991; Traxler *et al.*, 1991; Prashad, 1993; Barber *et al.*, 1999; Andres *et al.*, 2000; El Zoieby *et al.*, 2003). While this simulated docking has provided additional insights into structure–activity relationships, crystal structures of MTB RmlC complexes will be needed to validate these molecular interactions. Structure-guided *in silico* techniques have been used to rationalize and prioritize compound library design, but not all compounds showing RmlC activity have demonstrated activity against *Mycobacterium tuberculosis* in culture (Ma *et al.*, 2001; Babaoglu *et al.*, 2003). The permeability barrier of the mycobacterial envelope and new routes for drug delivery must also be considered, exploiting functionalities known to exhibit 'good' deliverability in library design and optimization of lead structures.

4. Conclusions

Although the architectures of the active sites of the RmlC enzymes from *Mycobacterium tuberculosis*, *Streptococcus suis*, *Salmonella typhimurium* and *Methanobacterium thermoautotrophicum* are conserved, there are notable differences that emerge based on structural alignment and pocket characterization which contribute to variation in binding-pocket size and shape. In the absence of published crystal structures of MTB RmlC–ligand complexes, such differences in active-site pocket character have a significant impact on structure-guided drug-design approaches that exploit *in silico* protein-ligand docking and underscore the need for ligand-bound crystal structures that more accurately define the protein receptor for simulated docking analysis and elucidate structure–activity relationships.

We acknowledge Uhn Soo Cho, Susan Wachocki, Minyoung So and Min-Young Kim for technical assistance with cloning and protein expression and purification at the LANL TB consortium protein-production facility, Fu Ming Tao at CSU Fullerton for helpful discussions and Clare Smith, Texas A&M University, for critical review of the manuscript. KAK thanks the California State University Program for Education and Research in Biotechnology and the W. M. Keck Foundation for support of the Center for Molecular Structure. LLNL is operated by University of California for the US DOE under contract W-7405-ENG-48. This work was funded by NIH P50 GM62410 (TB Structural Genomics) center grant. AZ was supported by LLNL LDRD grant 02-LW-003.

References

Abagyan, R., Batalov, S., Cardozo, T., Totrov, M., Webber, J. & Zhou, Y. (1997). *Proteins*, **29**, 29–37.
Abagyan, R., Totrov, M. & Kuznetsov, D. (1994). *J. Comput. Chem.* **15**, 488–506.
Andres, C. J., Bronson, J. J., D'Andrea, S. V., Deshpande, M. S., Falk, P. J., Grant-Young, K. A., Harte, W. E., Ho, H. T., Misco, P. F., Robertson, J. G., Stock, D., Sun, Y. & Walsh, A. W. (2000). *Bioorg. Med. Chem. Lett.* **10**, 715–717.
Babaoglu, K., Page, M. A., Jones, V. C., McNeil, M. R., Dong, C., Naismith, J. H. & Lee, R. E. (2003). *Bioorg. Med. Chem. Lett.* **13**, 3227–3230.
Barber, A. M., Hardcastle, I. R., Rowlands, M. G., Nutley, B. P., Marriott, J. H. & Jarman, M. (1999). *Bioorg. Med. Chem. Lett.* **9**, 623–626.
Becke, A. D. (1988). *Phys. Rev. A*, **38**, 3098–3100.
Biller, S. A., Forster, C., Gordon, E. M., Harrity, T., Rich, L. C., Maretta, J. & Ciosek, C. P. (1991). *J. Med. Chem.* **34**, 1912–1914.
Canutescu, A. A., Shelenkov, A. A. & Dunbrack, R. L. Jr (2003). *Protein Sci.* **12**, 2001–2014.
Christendat, D., Saridakis, V., Dharamsi, A., Bochkarev, A., Pai, E., Arrowsmith, C. H. & Edwards, A. E. (2000). *J. Biol. Chem.* **275**, 24608–24612.
Cruickshank, D. W. J. (1999). *Acta Cryst.* D**55**, 583–601.
El Zoieby, A., Sanschagrin, F. & Levesque, R. C. (2003). *Mol. Microbiol.* **47**, 1–12.
Engh, R. A. & Huber, R. (1991). *Acta Cryst.* A**47**, 392–400.
Giraud, M.-F., Leonard, G. A., Field, R. A., Berlind, C. & Naismith, J. H. (2000). *Nature Struct. Biol.* **7**, 398–402.
Helm, D. van der, Lessor, A. D. Jr & Merritt, L. L. Jr (1962). *Acta Cryst.* **15**, 1227–1232.
Hooft, R. W. W., Vriend, G., Sander, C. & Abola, E. E. (1996). *Nature (London)*, **381**, 272–272.
Kantardjieff, K. A. & Rupp, B. (2003). *Protein Sci.* **12**, 1865–1871.
Kantardjieff, K. A. & Rupp, B. (2004). In the press.
Kissinger, C. R., Gelhaar, D. K. & Fogel, D. B. (1999). *Acta Cryst.* D**55**, 484–491.
Laskowski, R. A. (2001). *Nucleic Acids. Res.* **29**, 221–222.
Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *J. Appl. Cryst.* **26**, 283–291.
Lee, C., Yang, W. & Parr, R. G. (1988). *Phys. Rev. B*, **37**, 785–789.
Lipinksi, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. (2001). *Adv. Drug. Deliv. Rev.* **46**, 3–26.
Ma, Y., Stern, R. J., Scherman, M. S., Vissa,V. D., Yan, W., Jones, V. C., Zhang, F., Franzblau, S. G., Lewis, W. H. & McNeil, M. R. (2001). *Antimicrob. Agents Chemother.* **45**, 1407–1416.
McRee, D. E. (1999). *J. Struct. Biol.* **125**, 156–165.
Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.
Merritt, E. A. & Bacon, D. J. (1997). *Methods Enzymol.* **277**, 505–524.
Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.
Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **267**, 307–326.
Prashad, M. (1993). *Bioorg. Med. Chem. Lett.* **3**, 2051–2054.
Ramaswamy, S. & Musser, J. M. (1998). *Tuber. Lung Dis.* **79**, 3–29.
Reddy, V., Swanson, S., Sacchettini, J. C., Kantardjieff, K. A., Segelke, B. & Rupp, B. (2003). *Acta Cryst.* D**59**, 2200–2210.

Rupp, B., Segelke, B. W., Krupka, H. I., Lekin, T. P., Schafer, J., Zemla, A., Toppani, D., Snell, G. & Earnest, T. E. (2002). *Acta Cryst.* D**58**, 1514–1518.
Schapira, M., Abagyan, R. & Totrov, M. (2003). *J. Med. Chem.* **46**, 3045–3059.
Schroeder, E. K., de Souza, O. N., Santos, D. S., Blanchard, J. S. & Basso, L. A. (2002). *Curr. Pharm. Des.* **3**, 197–225.
Segelke, B. W. (2001). *J. Cryst. Growth*, **232**, 553–562.
Stern, R., Lee, T., Lee, T., Yan, W., Scherman, M., Vissa, V., Kim, S., Wanner, B. & McNeil, M. (1999). *Microbiology*, **145**, 663–671.
Terwilliger, T. C. *et al.* (2003). *Tuberculosis*, **83**, 223–249.
Totrov, M. & Abagyan, R. (2001). *The Thermodynamics of the Drug–Receptor Interaction*, edited by R. B. Raffa, pp. 603–624. New York: John Wiley & Sons.
Traxler, P. M., Wacker, O., Bach, H. L., Geissler, J. F., Kump, W., Meyer, T., Regenass, U., Roesel, J. L. & Lydon, N. (1991). *J. Med. Chem.* **34**, 2328–2337.
Weiss, M. S. & Hilgenfeld, R. (1999). *Biopolymers*, **50**, 536–544.
Zemla, A. (2003). *Nucleic Acids. Res.* **31**, 3370–3374.